SPANISH NATIONAL CANCER RESEARCH CENTER

FACULTY OF STATISTICAL STUDIES

BIOSTATISTICS MASTER

Master's thesis

Study of the distribution and behaviour of the "0" values in large *omic* data arrays.

Author: Helena Fidalgo Gómez

Co-tutors: Dra. Núria Malats Riera

Dra. Mª Teresa Pérez Pérez

María Dolores Alonso Guirado

SPANISH NATIONAL CANCER RESEARCH CENTRE

Madrid, 2020

SPANISH NATIONAL CANCER RESEARCH CENTER

FACULTY OF STATISTICAL STUDIES

BIOSTATISTICS MASTER

Master's thesis

Study of the distribution and behaviour of the "0" values in large *omic* data arrays.

Author: Helena Fidalgo Gómez

Co-tutors: Dra. Núria Malats Riera

Dra. Mª Teresa Pérez Pérez

María Dolores Alonso Guirado

Dra. Núria Malats Riera | Dra. Mª Teresa Pérez Pérez

Helena Fidalgo Gómez | María Dolores Alonso Guirado

Madrid, 2020

"In the middle of difficulty, lies opportunity"

Albert Einstein (1879 – 1955)

Index:

1. Synopsis:

There is evidence that many diseases are not only determined by gene alterations. A clear example is cancer, encompassing several complex diseases where both genetic and non-genetic factors interact over the lifespan; the latter including environment exposures and the microbiome that can be assessed using *omics* approaches. *Omics* is a recent area of study including several biological disciplines. The technologies applied to *omic* sciences allow the study, at a molecular level, of the different elements that make up biological systems. Recently, biomedical science focusses on a new area: microbiome, where various associations between certain microorganisms and diseases have been found. One of the challenges in modelling cancer risk is the analysis of microbiome data: microbe counts are sparse and the data are high dimension and contain a large proportion of zeros.

This project aims to show different alternatives for the analysis of counting data that are characterized by a clear over-dispersion and excess of zeros. Moving ahead from classical linear models, there are regression models, such as Zero Inflated models or *"Hurdle"* models, I was able to establish what kind of zeros are in the database. These models and their corresponding distributions are subjected to different selection criteria with the purpose of establishing which is the model that best fits the data, depending on the percentage of zeros that it presents. By applying these approaches, I could appropriately define relationships between different microorganisms and gene expressions, tumour stages, immune subtypes, gender, BMI, ...

Keywords: microbiome, overdispersion, zeros, regression models, *Hurdle* models.

Existen evidencias de que muchas enfermedades no están determinadas sólo por alteraciones genéticas. Un claro ejemplo es el cáncer que engloba muchas enfermedades producidas por la interacción de factores genéticos y no-genéticos durante toda la vida. Entre los factores no-genéticos se encuentran la forma en que los seres humanos viven e interactúan con el medio ambiente y el microbioma; ambas exposiciones pueden ser caracterizadas con datos *ómicos*. Las tecnologías *ómicas* representan una reciente área de estudio que engloba diversas disciplinas biológicas. Las tecnologías aplicadas a las *ómicas* permiten estudiar, a nivel molecular los diferentes elementos que componen los sistemas biológicos. Hoy en día, el foco se encuentra en una nueva área: la microbioma, puesto que se han encontrado diversas asociaciones entre ciertos microorganismos y enfermedades. El reto principal en el análisis de datos de microbioma es el escaso número en los datos de conteo de microbioma, los cuales son de gran dimensión y contienen una gran proporción de ceros.

En este proyecto se pretende mostrar diferentes alternativas para el análisis de datos de conteo que se caracterizan por una clara sobredispersión y exceso de ceros. Aplicando modelos de regresión como los modelos de inflación de cero o los modelos *Hurdle* pude establecer qué tipo de ceros se encuentran en la base de datos. Estos modelos y sus correspondientes distribuciones están sometidos a diferentes criterios de selección con el objetivo de establecer cuál es el modelo que mejor se ajuste a los datos en función del porcentaje de ceros que presente. Ello me ha permitido definir relaciones entre diferentes microorganismos y expresiones genéticas, estadios tumorales, subtipos inmunes, género, IMC, ...

Palabras clave: microbioma, sobredispersión, ceros, modelos de regresión, modelos *Hurdle*.

2. Introduction to *Omics* Data:

Omics is comparatively new space of study that cut across the biological disciplines and has relevancy to all biological sciences. It attempts to look at biological systems in a holistic way and to account for all interactions between genes, proteins, RNA and metabolites (Arivaradarajan & Misra, 2019). From the second half of the XX century, and mainly at the end of it, technologies based on the progress of molecular biology were developed. In their development, these technologies were grouped around what is known globally today as the "*omics sciences*" and which refer to the knowledge derived from the application of a set of technologies that make possible the study at a molecular level of the different elements that make up biological systems in all their complexity, including the result of the interactions and relationships that occur between the internal components of the individual and the external elements. The technologies used by *omics*, called "*omic technologies*", are mainly characterized by generating massive amounts of data (also called big data) in a single experiment from a single sample.

It was as a result of the Human Genome Project in 1990 that genomics emerged as the first *"omics*" to be studied, which was based on the study of the genome or DNA. At the beginning of the XXI century the sequence of the human genome was reported (Venter et al., 2001), the order of all the nucleotides contained in the human DNA was deciphered *(International Human Genome Sequencing Consortium, 2004).* Although genomics offers correlations between diseases and gene variants, it does not demonstrate how that variable can be a cause of the disease. Therefore, it is necessary to integrate these correlations with other *"omics*" to find the function of genes and the variants that modulate them, in order to better understand the cause of a certain disease.

Thanks to the central dogma of biology, it is known that DNA is transcribed into messenger RNA (mRNA), which is also known as transcript *(Figure 1)*. Transcriptomics is the *"omics*" that studies the expression of transcripts coming from different genes. The mRNA is specific to each cell and to the pathophysiological conditions at a given time; thanks to the methodologies used to analyze the mRNA, comparisons can be made between cases and/or controls, or to check which genes are expressed under certain conditions. Currently, one of the applications of transcriptomics is the analysis of the expression of genes involved in different types of cancers. (Unger-Saldaña et al., 2015). The mRNA is translated into proteins, which are made up of aminoacids, which are responsible for performing the corresponding function of the gene. Proteomics studies thousands of proteins present in a sample. Once the proteins are translated, they can undergo post-translational modifications, these modifications cause structural changes that control the formation of functional protein complexes: they regulate the activity of the proteins and transform them into active or inactive forms. The Human Proteome Project has provided a list of proteins that have been found in different cell types and organs from different people. This data is public and reports more than 30,000 identified proteins in humans (Omenn et al., 2015).

Metabolites are those molecules that participate as substrates, intermediaries or products in the chemical reactions of metabolism; it is defined as a technology for determining the overall changes in concentration of metabolites present in a fluid, tissue or organism in response to a genetic variation, physiological or pathological stimulus (Cambiaghi et al., 2017). Metabolomics allows us to analyze the metabolic profile of a sample, both quantitatively and qualitatively.

In recent decades it has been shown that DNA can fold into three-dimensional structures that can regulate distant regions. Therefore, the sequence of nucleotides, then, is not the only thing that regulates gene expression, but how "tangled" the DNA is and its positioning during the formation of complex structures, that make up chromosomes. Epigenetics refers to the set of processes by which gene transcription is regulated without affecting the DNA sequence, this science represents the overall epigenetic changes in a sample, at a given time and under specific pathophysiological conditions.

There are as many *omics* sciences as biological or molecular elements that can be studied by these technologies, not only are sciences such as proteomic, transcriptomic or genomic taken into account; the list has been extended in recent decades, and there are currently research groups focused on these new areas of knowledge, which in the future will be considered as "emerging" *omics field*, including nutrigenomics, exposomics, lipidomics or metagenomics.

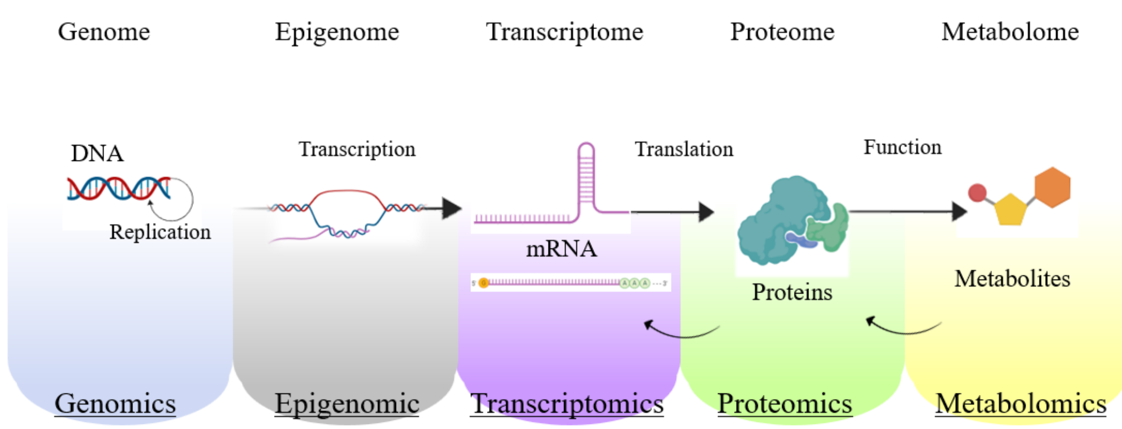


***Figure 1**. The central dogma of molecular biology and its relationship to the omics sciences: the genetic information contained in the DNA is transcribed into messenger RNA (mRNA) and non-coding RNA. The mRNA gives rise to proteins, through a process called translation. Proteins carry out the cellular functions, through which they generate or consume different metabolites. These metabolites and proteins, together with non-coding RNA, are involved in the regulation of these processes, forming a highly complex system of interactions. In this scheme, we summarize graphically what molecular aspect each of the "established" omics field encompasses in relation to the internal elements of the individual. Image created with Biorender.com.*

All these *omics* have a crucial role in the health sciences: they help to search for biological markers of diagnosis and prognosis of diseases, discover therapeutic targets against which new drugs could be developed, determine which molecular mechanisms are involved in the pathogenesis of diseases, among others. The amount of information obtained with these techniques is such that it exceeds human discernment, which is why powerful statistical and computer techniques are needed to interpret the data obtained; bioinformatics and biostatistics become essential tools behind the development of any *omics* and necessary for the management of them.

Today the focus is on one of these "emerging" *omics*, metagenomics, where various associations between certain microorganisms and diseases are being found. In recent years, taking the Human Genome Project as a base, several projects related to this field have been developed, the Human Microbiome Project (USA) (Turnbaugh et al., 2007), (Hutchison, 2007), and the European Project MetaHIT (*MetaHIT Consortium -Metagenomics of the Human Intestinal Tract Consortium-Wellcome Sanger Institute*), with Spanish contribution. The latter was selected by the prestigious scientific journal *Science* as one of the 10 most notable discoveries of 2011. Both are responsible for studying the microorganisms that inhabit our bodies through metagenomics, which analyzes the genome of all the microorganisms in a population as a whole and how they can react to different stimuli.

2.1. Microbiome and Metagenome:

Ten years once the term metagenomics was coined, the approach continues to collect momentum. Metagenomics is the study of genetic material of environmental samples of cohabit. This scientific discipline makes it possible to sample the genome sequences of a community of organisms living in a common environment (Hugenholtz & Tyson, 2008).

There is a wide variety of microbial communities and their genes, the microbiome, throughout the human body, with fundamental functions in human health and disease. In 2008, the National Institutes of Health of the United States (NIH) promoted a study with a five-year forecast under the name of *Human Microbiome Project* (HMP), where the effects of microbes, as well as viruses, bacteria and microorganisms, and how they influence the state of human health have been addressed. Where it was concluded that the human body contains more microbes than human cells (Group et al., 2009). The objective of HMP is to explain the microbe communities found in numerous elements of the human body and to obtain correlations between changes within the microbiome and also the health of individuals. Therefore, the bacteria found in the intestinal microbiota are a key part of HMP research, the composition of our intestinal bacteria affects the maturation of the human immune system and is a relevant factor in the development of not only gastrointestinal but also cardiovascular diseases. Its links with cancer and diabetes are under active investigation (Requena & Velasco, 2019)

The diversity of microbes within a given habitat of the body can be defined as the distribution of the number and abundance of different types of organisms, which has been linked to various human diseases. It is estimated that there are between 500 and 1,000 species of bacteria throughout the human body, and each bacterial strain has a genome containing thousands of genes, offering substantially more genetic diversity and flexibility than the human genome. However, different people host completely different collections of microbes with densities that vary substantially, even among conserved taxa, and little is understood about what leads to variation and what regulates it *(Graph 2)*. Importantly, research still does not know how microbial variation within a person over time or among different people influences well-being, the preservation of health, or the onset and progression of disease (Gilbert et al., 2018).

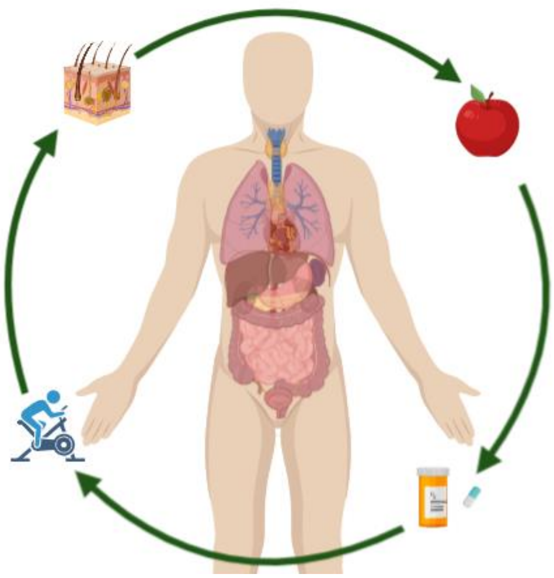


Figure 2. *The human microbiome (gut, skin, bladder, and vagina, among other organs) and the factors that influence it: the human microbiome follows specific trajectories of the body site, so that each organism develops a specific biogeography. Excessive cleaning can temporarily alter the skin microbiome; changes in diet can profoundly affect the structure of the intestinal microbial community. The effect of antibiotics on all microbiomes is expected to be large relative to that of other factors, and preliminary studies have been conducted to determine their impact. It is also believed that lifestyle has a strong influence on the composition of the microbiome, for example exercise seems to influence the structure of the microbiome through the reduction of inflammation, resulting in subtle changes in the composition of the microbial community. Image created with Biorender.com*

In recent years, a great deal of evidence has been generated, reinforcing the importance of the human microbiome in health and disease through various mechanisms. The host microbiome also supplies a physical threshold that protects the host from extraneous micro-organisms, through competitive suppression and the development of antimicrobial agents (*The Human Microbiota in Health and Disease | Elsevier Enhanced Reader*).

Early microbial exposure plays a critical role in the development of the immune, endocrine, metabolic and nervous systems (Requena & Velasco, 2019). In addition to the importance of adequate training of the immune system in the early stages of life, recent studies of microbiota alterations (*dysbiosis*) and their association with certain pathological conditions point to the decline in microbial diversity as one of the aspects contributing to the development of diseases, the reduction in species diversity of the human microbiome has been associated with an increase in pathologies (Sommer et al., 2017). Many autoimmune, allergic, endocrine-metabolic and gastrointestinal diseases are associated with microbial imbalance, making it a current issue in medicine and biology.

There is evidence that most diseases are not determined by genes; a clear example is cancer. The American Cancer Society in the United States has analyzed several genes associated with cancer in the last 50 years and there was only a 5% correlation between these diseases, globally, and certain genes. Therefore, the remaining 95% of cancer patients do not have a strictly genetic cause, but rather produced by the way human beings live and interact with the environment or the microbiome, with a host-microbiome lifespan interaction (Ariza-Andraca & García-Ronquillo, 2016) .

The most prominent examples are *Helicobacter pylori*, which is involved in the development of gastric cancer, and the high-risk types of human papillomavirus in cervical cancer. The interaction of microorganisms and their hosts is extremely complex, and a multitude of molecular mechanisms can be predicted by which they influence oncogenesis, tumor progression and response to cancer treatment (Garrett, 2015).

2.1.1. The Urinary Microbiome:

In 2008, when the results of the Human Microbiome Project were published, the microbiome of the urinary tract was not considered due to concerns about sampling techniques, microbiological methods and the concept of the urinary tract as a sterile niche. Thanks to new advances in advanced molecular techniques for the analysis of urine, the dogma that urine is sterile has been overturned and dysbiosis of the urinary microbiota has been linked to urological disorders. In view of this fact, the first results began to provide new insights into functional urological disorders (Aragón et al., 2018). Advances in 16S rRNA sequencing technology made it possible to discover the presence of a rich and diverse urinary microbiota in every individual, up to 80% of bacteria can be isolated using modified culture techniques. These novel technical approaches were used to investigate the composition of the urinary microbiota in patients diagnosed with functional disorders such as interstitial cystitis, urgency urinary incontinence, pelvic pain syndrome and bladder cancer.

Bladder cancer is considered the ninth most common malignant disease, with over 160,000 deaths recorded worldwide each year. The risk of developing this tumor grows with age *(Figure 3)*, and it is detected three times higher in men than in women (Sanli et al., 2017). In addition to environmental and genetic risk factors, researchers are increasingly aware that microbes in the human body play an important role in the maintenance of health and the development of disease *(Bladder Cancer Risk Factors, American Cancer Society, 2019).*

Over the past five years, evidence has been collected that the urinary tract is home to a variety of commensal microorganisms, and the urinary microbiome reported for healthy individuals varies considerably due to the use of different analytical and sample collection methods. In general, it can be stated that there are differences related to sex and age, as well as significant inter-individual variability in the composition of the urinary microbiome (Aragón et al., 2018).

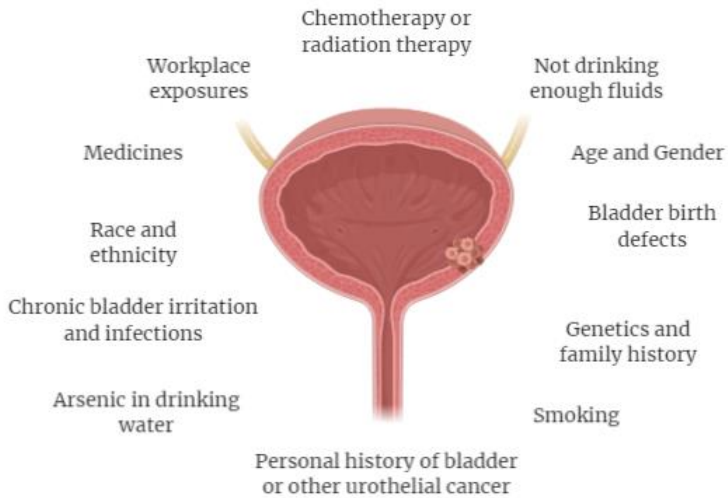


Figure 3. *Bladder Cancer Risk Factors: According to the American Cancer Society, a risk factor is anything that affects your chance of getting a disease like cancer. Different cancers have different risk factors, in the case of bladder cancer there are risk factors that we can control such as lifestyle, not smoking, drinking water…, other factors such as race or genetics and family history, cannot be changed. Having more risk factors will make a person more likely to present a bladder cancer. The overall 5-year survival rate for people with bladder cancer is 77%. Image created with Biorender.com*

The urinary microbiome in bladder cancer has barely been investigated, the Xu et al. in 2015, pilot study reporting enrichment of *Streptococcus sp.* in some of the cancer patients. In 2018, in a study conducted by Bučević Popović et al., they analyzed urine samples from a total of 23 subjects: 12 individuals with bladder cancer and 11 healthy controls, using 16S sequencing, showing no significant differences with respect to microbial diversity or overall microbiome composition. The authors did note, however, that *Fusobacterium*, a genus associated with colorectal cancer, was enriched in the urothelial cell carcinoma group. In the same year, Wu et al., conducted a study to characterize the urinary microbial community associated with bladder cancer; they collected urine samples from 31 male patients with bladder cancer and from 18 non-neoplastic controls using 16S sequencing. They identified an increased abundance of bacteria such as *Herbaspirillum*, *Porphyrobacter* and *Bacteroides* in cancer subjects with a high level of relapse and gradation, indicating that these types of OTUs are biomarkers for high-risk bladder cancer.

In 2019 another study was carried out by Bi et al., where the same methodology was used; the urine samples of 29 patients with bladder cancer and 26 controls were compared. They discovered that the urine samples contained a microbiota preserved from *Firmicutes*, *Actinobacteria*, *Proteobacteria* and *Bacteroidetes*, constituting 94.4% of the bacteria in every situation; but only people with bladder cancer had a greater presence of *Actinomyces europaeus*, which may indicate that this strain is representative of bladder cancer (Bi et al., 2019).

During this last year, the American Society of Clinical Oncology, conducted a study to characterize the urinary microbiota associated with muscle and muscle invasive bladder cancer; they collected urine samples from 27 patients with bladder cancer prior to transurethral resection or cystectomy and from 12 non-neoplastic control subjects based on age. The species most predominant in the groups were *Proteobacteria, Bacteroidetes, Firmicutes* and *Actinobacteria*, but *Bacteroidetes* being a little more abundant in patients with bladder cancer. Interestingly, they found that non-muscle invasive bladder cancer displayed a reduction in the abundance of *Sphingobacteriaceae*, *Bifidobacteriaceae* and *Enterobacteriaceae*. The study concluded that the urinary microbiota of patients with bladder cancer shows a totally different pattern compared to the lipid control, so the microenvironment of the tumor may influence the dysbiosis (Oresta et al.,2020).

All these studies come to the same conclusion: the urinary microbiota may be a potential biomarker and therapeutic target for bladder cancer. But in order to identify which bacterial taxa are associated with this pathology it is necessary to carry out the correct statistical analyses; microbiome research and data analysis is one of the fastest growing sectors in biomedical and public health research, with an enhancement in the number of microbiome-funded studies and publications about statistical methodology on microbiome data. Because microbiome data is so complex, there is a critical need to develop all types of statistical methodologies for microbiome research, ranging from application to methodology to statistical theory (Xia et al., 2018).

3. Microbiome Data:

Microbiome data is generated, frequently, through 16S rRNA gene sequencing and shotgun metagenomics sequencing. Once the raw sequences have been pre-processed, there are two ways to generate microbiome data for further analysis (J. Chen et al., 2013). These 16S sequences are grouped into Operational Taxonomic Units (OTUs), according to their similarity (Caporaso et al., 2010).

Depending on the fields of research and the bioinformatics tools used to obtain this type of data, the structure of the microbiome and/or genome data is usually similar to contingency table per sample. This type of table usually has the samples as columns and the traits as rows. In general, the features refer to any of the characteristics of the OTUs and the "samples" are also items, i.e. the rows of the data array can be items, while the columns are variable. In the literature on microbiomes, researchers often use OTUs, taxa, genera and species to refer to characteristics. Thus, the primary data structure in the study of microbiomes is a table of taxa. Therefore, when working with microbiome data, these structures are referred to as taxa per sample or sample per table (Xia et al., 2018).

Microbiome data have several features. Microbe count data (OTU counts) are limited, high dimension, poor and contain an excess of zero counts. In the OTU matrix, there are complicated correlation structures between various taxa, displaying a clear over-dispersion with great heterogeneities inside the group. Specifically, the OTU table records counts of different bacteria in samples extracted from the 16S rRNA sequencing. Each row of the table corresponds to a genus while each column records a read count corresponding to a sample. The levels of phylum, class, family, order and species have the same data structures (Xia et al., 2018).

Microbe sequence data sets are large, with tens of thousands of different categories. They are underdetermined, presenting a much larger number of OTUs than the number of samples (Tsilimigras & Fodor, 2016). The high dimensionality could result in the *"large p, small n"* problem (Yin & Hilafu, 2015) and poses statistical challenges to analyse microbiome data. High dimensionality is not the only problem with these data, taxon count data, whether taxonomy readings or OTUs counts, are often widely dispersed. This suggests that the variance of the counts is higher than that predicted by a multinomial regression (e.g. with the Poisson distribution). The over-dispersion in the data is due to the size of the library of DNA or RNA sequencing are widely different, and to the OTUs count ratios, because they differ much more than was expected under the multinomial regression proposed, like the Poisson model. (McMurdie & Holmes, 2014).

In microbial data, dispersion is perceived as the absence of many taxa in the samples and zeros are generated in most experiments. The abundance of microbiome taxa, especially the abundance of taxa at lower taxonomic levels or OTU counts, are often high and highly skewed (Xia & Sun, 2017).

3.1. Zero inflation challenge:

Microbiome count data are represented using OTUs from the 16S rRNA studies. For each specimen taken from a certain ecosystem, the number of occurrences of each OTU is measured and the resulting table of OTUs is summarized to obtain the relative abundance for the bacterial taxa in a specimen. An important feature of these data is that not all taxa can be present in each sample, i.e. some of the OTUs can take zero values. There is an urgent demand of statistical methods to analyse these complex microbial count data. This is an active area of research and a variety of statistical and computational methods have been proposed in the literature to answer a variety of questions.

The excess of zeros in microbiome count data demonstrates a challenge in analyzing this type of data, especially when comparing two or more groups. A common strategy for managing these excess zeros is to use various probability models to model the excess zeros (Paulson et al., 2013; E. Z. Chen & Li, 2016). The problem with working with this type of model is that, in many cases, an implicit assumption is made that all zeros can be explained under a common probability model. In the count data modeling, three kinds of zeros are often referred based on the sources of zeros (Kaul et al., 2017):

- *Outlier Zeros*: this kind of zeros are declared as outliers by the methodology employed; a taxon is recorded as outlier zero due to extraneous reasons, but not because it is below the detection limits due to the depth of the sample (Kaul et al., 2017). This definition has been found in only one article, Kaul et al. are the only ones that consider this type of zeros.

- *Structural Zeros*: in many occasions, due to the nature of the experimental groups, certain taxa are considered not to be present in the samples obtained from some groups, although they may be present in others (Kaul et al., 2017). A structural zero (Martín-Fernández et al., 2014), an essential zero (Aitchison & Kay, 2003; Martín-Fernández et al., 2014), genuine zeros, or the absolute zero (Martín-Fernández et al., 2014) refers to a given observation, when it is not correctly defined or simply cannot exist due to some deterministic reasons (van den Boogaart & Tolosana-Delgado, 2008). According to Aitchison and Kay in 2003, this type of zero means that "a component which is truly zero, not something recorded as zero simply because the experimental design or the measuring tool has not been able to detect a trace of the part".
  In summary, the zeros that truly represent the absence of taxa from a given sample belong to the structural zeros (Tsilimigras & Fodor, 2016).

- *Sampling or Counting Zeros*: count data are categorical data, in which the count represents the number of items falling into each of several categories (Martín-Fernández et al., 2014). If an observed zero in the data cannot be qualified as an outlier or structural zero, then this zero is declared to be a sampling zero, perhaps caused by the depth of the sampling (Kaul et al., 2017). According to Martín-Fernández et al., this type of zero is due to a sampling problem, since the components may not be observed due to the limited sample size or be undetectable due to the limit of the techniques, i.e., the zeros are due to insufficiently large samples (Martín-Fernández et al. 2014). Unobserved positive values may be observed with a larger number of tests or with a different sampling design. Kaul and others claim that this type of zero is due to sampling depth, potentially due to the fact that the taxon is relatively rare compared to other taxa and due to technological reasons was not observed.

Microbiome data have a large amount of zeros, either structural or sampling (e.g. biological and technical variability). Each individual has a unique composition of OTUs, because the taxa are dependent on the individual. The count of OTUs is, generally, characterized by an inflation of zeros (while one subject may have more than one hundred counts in a given OTU, many other subjects may have 0). (L. Xu et al., 2015).

Structural zeros of the taxa are perceived in a certain sample because they are biologically or physically absent in the sample or in the subject. Sampling zeros are due to the true discovery of low-abundance taxa that are only found in a few samples (Tsilimigras & Fodor, 2016).

3.2. Challenges of Modelling Microbiome Data:

Taxa or OTUs are distributed throughout the samples (subjects), in the form of integer numbers or counts, and are not usually distributed according to a normal distribution. Ordinary regression models, of which t-tests, ANOVA and ANCOVA are special cases, assume that the result is normally distributed and may produce a biased estimate of a treatment effect (and other factors) if that assumption is not met. This would imply, in practical terms, that the size of the treatment effect and its statistical significance are either overestimated or underestimated, which is not appropriate (Hu et al., 2011).

In the last decades, we have seen the increasing development and availability of statistical methods for parametric models whose data are not distributed according to a normal distribution, and which respond to the challenges currently facing biostatistics with microbial count data. These challenges are (Xia et al., 2018):

i. Reduce dimensions and solve *large p* and *small n* problem.
ii. Identify and manage "rare" taxa.
iii. Model the microbiome data with over-dispersion and inflation of zeros.

Over-dispersion is the main issue in the analysis of 16S rRNA sequence data (Tsilimigras & Fodor, 2016), so handling over-dispersion with excess zeros is the key issue in the analysis of microbiome data.

Over-dispersion with excess zeros poses critical challenges in parametric models in order to calculate accurate estimates of variance for meaningful inference and even such estimates are essentially impossible on samples that consist mostly of zeros (Tsilimigras & Fodor, 2016). When taxa or OTUs are over-dispersed with excess zeros, the abundance distribution and the probability of occurrence distribution of the taxa are both skewed, giving rise to zero inflation (J. Chen et al., 2013).

Classical linear models that apply to untransformed counts or log transformed counts are inappropriate for zero inflated count data, as they would violate the normality and homogeneity of variance assumptions, and are not relevant for relative abundance either (L. Xu et al., 2015). An example would be relative abundances, which are limited to between zero and one and where the variance is usually dependent on the mean. In addition, no data transformation can satisfy the assumptions if there is an excess of zeros; logistic regression, which considers all zero counts as "non-events", is generally used for modelling data with excess zeros; the problem with applying this method is the loss of information and power to detect the effect of a covariate (L. Xu et al., 2015). When faced with generalized linear models, such as Poisson's or the Binomial Negative model, which can be applied to count data (since they work with discrete and non-negative data), they cannot deal with the excess of zeros either, because a basic requirement of these models is that the proportion of zeros must necessarily be linked to the distribution of positive values (Xia et al., 2018). Thus, the abundance of OTUs with excess zeros cannot be modelled by any standard parametric model (Martin et al., 2005).

Over-dispersion with zero inflation also does not allow the application of non-parametric methods. If non-parametric methods that do not assume the assumption of normality are applied, such as the Wilcoxon rank sum test, covariates could not be incorporated, and a large loss of power would be assumed due to the large number of ties caused by many zeros. In addition, these methods are based on ranges or medians, so they tend to be more "robust" to outliers and prevent variance estimates that may be biased by small samples (Martín-Fernández et al., 2014). On occasions when many OTUs have an excess of zeros and few positive count values, they will not have sufficient statistical power to be able to make an inference about these types of taxa, using non-parametric methods. (Xia et al., 2018) .

Overall, both traditional parametric models and non-parametric methods are not suitable for analysing data on microbiomes with over-dispersion and excess zeros. Therefore, the analysis of this type of data is a real challenge and if excess zeros are not taken into account, biased parameter estimates, and misleading inferences can occur (Xia & Sun, 2017; Xia et al., 2018; L. Xu et al., 2015).

4. Statistical Methods for Microbiome Data:

For each specimen taken from an ecosystem, the number of occurrences of each OTU is measured and the resulting table of OTU is summarized to obtain the relative abundance of bacterial taxa in a specimen. These OTU counts can be summarized at any level of the bacterial phylogeny, e.g. species, genus, family, order, etc. (Kaul et al., 2017). If, for example, a group comparison is desired, there are classical methods that can be used: depending on whether the data are normally distributed or not, on the number of experimental groups or on the experimental conditions, a t-Student, an analysis of variance (ANOVA) or a corresponding non-parametric test can be applied. For example, t-test was used to compare alpha diversity (La Rosa et al., 2015) between two sets of relative abundance data. The non-parametric analogous Wilcoxon rank sum test was conducted to compare alpha diversity, like, Shannon diversity (La Rosa et al. 2015). Wilcoxon rank sum test was also used to identify the differences in OTUs and the relative abundances of different phyla and genera (Wang et al., 2012). When comparing more than two groups, the ANOVA test or its non-parametric equivalent, the Kruskal-Wallis test, is used. The Chi-Square test is generally used to compare categorical microbiome data, for example, testing if a single a priori specified taxon is present at different rates across groups (La Rosa et al. 2015).

In this type of situation, the classic methods do work, but when the objective is to model this type of data, it is necessary to propose different methods. The abundance of the taxa in human samples is characterised by a greater number of zeros at lower taxonomic levels. In order to model the excess of zeros, over-dispersion and heterogeneity presented by microbiome count data, it is necessary to apply models that move away from classical linear models, such as the Zero-Inflated model or the Zero-Hurdle model. (L. Xu et al., 2015).

Before defining these models, it is necessary to clarify the excess of zeros and the clear over-dispersion of microbial data. *Figure 4* shows an example of a real case of *Finegoldia magna*, that is part of the database that will be used in this project and will be shown later how to solve it:

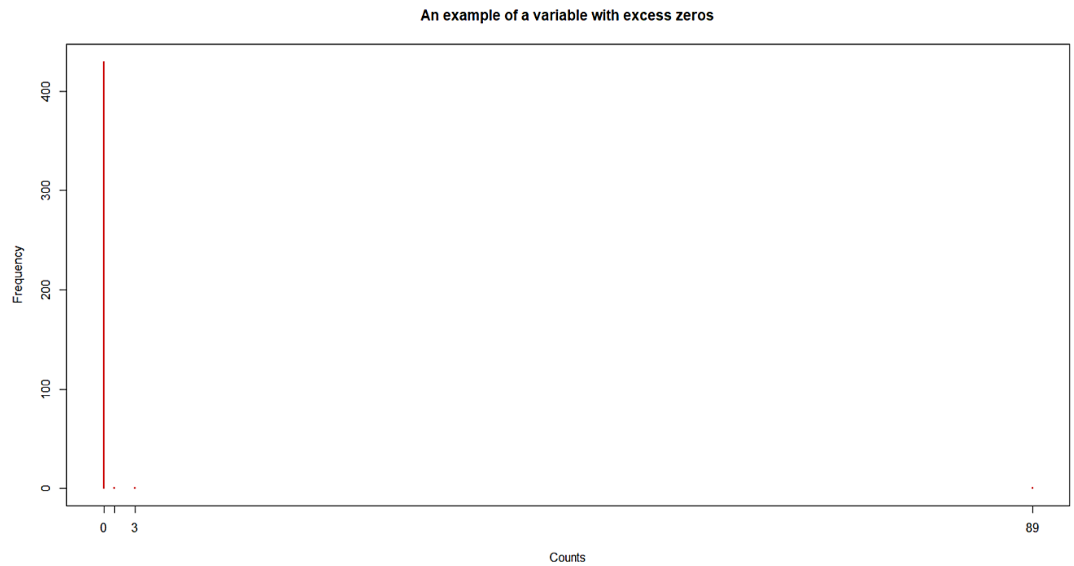


Figure 4. *Frequency of Count of an OTU with excess of zeros. The Y-axis shows the 433 subjects that make up the database and the X-axis shows the counts of each of these subjects. The graph above shows the frequency of "counts" for a specific bacterium (taken at random), as can be seen, out of 433 subjects in total: 430 subjects have 0 counts in this OTU, two subjects have less than or equal to 3 counts and one subject has 89 counts. Therefore, the excess of zeros is very large, and the variance far exceeds the expected value* ($Var(OTU_i) = 18.28 > E(OTU_i) = 0.22$)*, confirming that present overdispersion. Graph created in RStudio with the "ggplot2" package.*

Count data is discrete data, non-negative integer ($\mathbb{Z} \geq 0$), and are modelled according to a $\mathbb{P}$oisson or a $\mathbb{N}$egative $\mathbb{B}$inomial distribution. If a $\mathbb{G}$aussian distribution were used, a least-squares regression would be carried out, where the count data would violate the distribution assumptions on which the normal model is based and produce statistically biased results. The central problem is that the standard model assumes that negative values are possible and that the variance of the variable being modelled is constant across its range of values. These assumptions are not possible for count data (Hilbe, 2017).

- $\mathbb{P}$oisson Distribution is the discrete probability distribution of the number of events occurring in a given time period, so it can take any integer value greater than or equal to zero. A random variable X follows a Poisson distribution of parameter $\lambda > 0$ and is denoted by $X \sim P(\lambda)$, if for $k \in \{0, 1, 2, 3, \dots\}$, the probability function is given by (Inouye et al., 2017):

$$P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda} \tag{1}$$

  Where λ indicates the average number of events in the given time period. A central criterion of the Poisson distribution is that mean and variance of the counts being modelled are identical (Inouye et al., 2017):

$$\mathbb{E}(X) = \lambda; \qquad \mathbb{V}ar(X) = \lambda \tag{2}$$

- ℕegative 𝔹inomial Distribution is a discrete probability distribution of the number of successes in a sequence of independent and identically distributed Bernoulli trials before a specified number of failures (denoted r) occurs. So, a random variable X follows a Negative Binomial distribution of parameters $r \in \mathbb{N}$ and $p \in (0,1)$, and is denoted by $X \sim NB(r, p)$, the probability function is given by (Fisher, 1941):

$$P(X = k) = \binom{r + k - 1}{r - 1} p^r (1 - p)^k \tag{3}$$

Where the mean and variance are defined by (Fisher, 1941):

$$\mathbb{E}(X) = \frac{r(1 - p)}{p}; \qquad \mathbb{V}ar(X) = \frac{r(1 - p)}{p^2} \tag{4}$$

As discussed in previous chapters, microbiome data are represented by an excess of zeros, which can be of three types, Outliers, Structural and Sampling, and depending on the origin of these zeros, they will fit one type of distribution better than another. If the data fit better to a ℙoisson distribution, it is because the zeros it contains are outliers and structural, whereas, if they fit better to a ℕegative 𝔹inomial distribution it is because it contains outliers, structural and sampling zeros. Therefore, sampling zeros are the ones that cause the over-dispersion that must be modelled with ℕegative 𝔹inomial distribution (Tang et al., 2018, Kaul et al., 2017, L. Xu et al., 2015, HE et al., 2014).

4.1. Models for count data with excess of zeros:

The main objective is to determine if the abundance of a certain OTU is related to some environmental or genetic factor; but currently there is no standard method to evaluate such relationships (L. Xu et al., 2015) .

As already mentioned in the Section 2.2, classical linear models, using non-transformed or logarithmic transformed counts are inappropriate for zero inflated counts due to the violation of normality and constant variance assumptions. In addition, no data transformation can meet the assumptions if there are too many zeros (L. Xu et al., 2015, Kaul et al., 2017, Xia & Sun, 2017).

If a logistic regression is applied, it would treat all zero counts as non-events, resulting in a loss of valuable information and power to detect the effect of a covariate. Generalized linear models such as ℙoisson or ℕegative 𝔹inomial can be applied on sequence counts and the logarithm of total sequence reads can be set as an offset; however, they could not model the excess of zero either, since a basic requirement of these models is that the proportion of zeros must be linked to the distribution of positive integer values (Xia et al., 2018, L. Xu et al., 2015). One way to handle excess zeros is to apply either a Zero-Inflated (ZI) model or a two-part model, also called the Zero-Hurdle model. With the Zero-Inflated models a better result will be obtained if outliers and structural zeros are being modelled, on the other hand, the hurdle models do not distinguish between the different types of zeros and model them perfectly, independently of their origin (L. Xu et al., 2015):

i. Zero Inflated Models:

This type of model is based on a distribution that allows for zero-valued observations, i.e. it is based on a zero inflation probability distribution (this is the name given to the regression model). The zero inflated models include Zero Inflated Poisson (ZIP) and Zero Inflated Negative Binomial (ZINB); they assume that for each observation, there are two possible data generation processes, and then a Bernoulli trial determines which process is used (L. Xu et al., 2015).

ZIP assumes that each observation comes from one of two potential distributions, one consisting of a constant zero while the other following a Poisson distribution. In this model, a logit model is typically used to analyse the probability of the outlier zero or structural zero, whereas the count data is analysed by the Poisson regression (Hu et al., 2011) .

Specifically, if the outcome variable for the j^{th} individual, Y_j, follows a ZIP distribution, then its probability is given by (Xia et al., 2018) :

$$P(Y_j = 0|X_j) = p_j + (1 - p_j) * e^{(-\mu_j)}; \tag{5}$$

$$P(Y_j = y_j|\, X_j) = (1 - p_j) * \frac{exp(-\mu_j)(\mu_j)^{y_j}}{y_j!};\ y_j = 1, 2, 3, \ldots \tag{6}$$

Where p_j is the probability of structural or outlier zero, $(1 - p_j)$ is the probability of sampling zero, μ_j and X_j are the expected Poisson count and covariate vector respectively, for the j^{th} individual.

From the first situation, it obtains that the observed zeros arise from both the zero-component distribution and the Poisson distribution. i.e., the two sources of structural and outliers' zeros, and sampling zeros. Therefore, the zero-component distribution provides the capability to model the "excess" zeros that are observed in addition to the zeros that are expected to be observed under the assumed Poisson distribution. By replacing the Poisson distribution for the count data in ZIP with the Negative Binomial distribution, it obtains the Zero-Inflated Negative Binomial distribution, or ZINB (Xia et al., 2018). Thus, a ZINB has the general form:

$$P(Y_j = 0|X_j) = p_j + (1 - p_j)g(\mu_j); \tag{7}$$

$$P(Y_j = y_j|\, X_j) = (1 - p_j)f(\mu_j)\ ;\ y_j = 1, 2, 3, \ldots \tag{8}$$

Where $g(\mu_j) = P(Y_j = 0 | X_j) = \left(\frac{\alpha^{-1}}{\mu_j+\alpha^{-1}}\right)^{1/\alpha}$ in the count data model, and $f(\mu_j)$ is the density of the Negative Binomial distribution; α is the dispersion parameter:

$$f(\mu_j) = \frac{\Gamma(y_j + \alpha^{-1})}{y_j!\,\Gamma(\alpha^{-1})} * \left(\frac{1}{1+\alpha\mu_j}\right)^{1/\alpha} * \left(\frac{\alpha\mu_j}{1+\alpha\mu_j}\right)^{y_j}; \;\; \alpha > 0, y_j \geq 0 \quad (9)$$

The binary process can be modeled using either *logit*, *probit* or other models for binary outcomes. For ZINB, $Var(Y_j| X_j) > E(Y_j| X_j)$, demonstrating that ZINB has the capability to model overdispersion. If we compare the ZIP and ZINB model, the main advantage of the ZINB is that the Binomial Negative distribution models the overdispersion and solves all sources of heterogeneity presented in the data. While the ZIP model, being based on a Poisson distribution, focuses on the heterogeneity created by the atypical and structural zeros. As Poisson is nested within Negative Binomial, ZIP is nested within ZINB; ZINB can be viewed as an extension of ZIP in analogous to Negative Binomial distribution being an extension of Poisson distribution (Hu et al., 2011).

ii. Two-part Models or Zero-Hurdle Models:

The Hurdle model was developed separately by Mullahy in 1986: "The idea underlying the hurdle formulations is that a binomial probability model governs the binary outcome whether a count variate has a zero or a positive realization (i.e., a transition stage). If the realization is positive the 'hurdle' is crossed, and the conditional distribution of the positives is governed by a truncated-at-zero count data model".

In short, one distribution addresses the zeros while another distribution addresses the positive nonzero counts, such as a truncated Poisson or truncated Negative Binomial distribution (Min & Agresti, 2005). It is a "finite mixture generated by combining the zeros generated by one density with the zeros and positives generated by a second zero-truncated density separately…" (Mullahy, 1986). A very important characteristic of the transition model is asymmetry, which means that the probability of crossing the obstacle increases as the covariates increase, and decreases as the covariates decrease.

Mullahy demonstrated that hurdle models naturally admit overdispersion and underdispersion of data, and because of this, give better results than Zero-Inflated models; given any two probability distribution functions for non-negative integers f_1 and f_2, presenting the hurdle part and the parent process, the hurdle-at-zero model has the probability distribution (Mullahy, 1986):

$$P(Y = 0) = f_1(0) \quad (10)$$

$$P(Y = y) = f_2(y) * \frac{1 - f_1(0)}{1 - f_2(0)} = \Phi f_2(y); \;\; y = 1,2,\ldots \quad (11)$$

Because $f_1(0)$ is essentially used to establish the event of crossing the hurdle (i.e., if the count is zero), is defined $P(Y = 0) = p_j$ and $P(Y \geq 1) = 1 - p_j$, and used a logistic regression to model p_j. The numerator of Φ can be interpreted as the probability of crossing the hurdle (in microbiome data would be interpreted as: in case of bacteria species (OTU), the probability to present at least one count) and the denominator is a normalization for f_2. It follows immediately that the hurdle model collapses to the parent model if $f_1 = f_2$, $\Phi = 1$ (Mullahy, 1986).

Since the zero hurdle model has two stages, a link function is applied for each stage:

- For the logistic regression:

$$\ln\left(\frac{p_j}{1 - p_j}\right) = X_j\alpha \tag{12}$$

- For the truncated model:

$$\ln(\mu_j) = X_j\beta \tag{13}$$

Where α and β are the sets of regression coefficients for the stages 1 and 2, respectively. One thing that needs to be stressed is that different prediction variables X could be used at each stage of the model, as the prediction variables that would be applied at the first stage may not be the same as those that would be applied at the second stage (this would be applied in multiple regression models).

Zero-Hurdle Poisson Model (ZHP) is a two-component model: a hurdle component models the zero versus the non-zero counts, and a truncated Poisson count component is employed for the non-zero counts (Xia et al., 2018):

$$P(Y_j |\, X_j) = p_j \tag{14}$$

$$P(Y_j = y_j |\, X_j) = (1 - p_j) * \frac{exp(-\mu_j)(\mu_j)^{y_j}}{y_j!\left(1 - exp(\mu_j)\right)};\ \ y_j > 0 \tag{15}$$

Specifically, zero-hurdle models do not make the distinction between outliers, structural and sampling zeros and handle them identically: unlike p_j in the zero-inflated model, the p_j in zero-hurdle model does not model the excess zeros, but all zeros (Xia et al., 2018).

The Zero Hurdle Negative Binomial model (ZHNB) is obtained by replacing the zero-truncated Poisson with a truncated Negative Binomial model to analyze the truncated-at-zero count:

$$P(Y_j |\, X_j) = p_j \tag{16}$$

$$P\left(Y_j = y_j \middle| X_j\right) = \left(1 - p_j\right) * \frac{\Gamma\left(y_j + \alpha^{-1}\right)}{\left(1 - \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_j}\right)^{1/\alpha}\right)\Gamma\left(y_j + 1\right)\Gamma\left(\alpha^{-1}\right)} * \left(\frac{\alpha^{-1}}{\alpha^{-1} + \mu_j}\right)^{1/\alpha} * \left(\frac{\mu_j}{\alpha^{-1} + \mu_j}\right)^{y_j} ; \quad for\ y_j > 0 \tag{17}$$

Where $\alpha(\geq 0)$ is a dispersion parameter that is assumed not to depend on the covariates. It can be seen in above equation, that the positive count is governed by a truncated-at-zero negative binomial as the probability function for the positive count is divided by 1 minus the probability function of a negative binomial evaluated at zero (Xia et al., 2018).

So far, it has been described and differentiated between Zero Inflated models and Zero Hurdle models from modelling and concepts. In general ZIP vs. ZHP, ZINB vs. ZHNB, respectively give similar model fit and predicted values, but different estimated parameters. These two models differ from their interpretation of model parameters, conceptualization of zeros and their capacity to deal with excess zeros. Zero-inflated models fit better if the zeros presented by the variable are outliers or structural; if it is analysed in the scope of these models, the following link functions can be specified for the count and the binomial data, respectively (x_j is used to indicate the covariate):

$$\mu_j = exp\left(\alpha + \beta * x_j\right) \tag{18}$$

$$p_j = \frac{exp\left(\upsilon + \gamma x_j\right)}{1 + exp\left(\upsilon + \gamma x_j\right)} \tag{19}$$

The mean μ_j for the Poisson count data is modeled in terms of the covariate x_j and the probability p_j for the Binomial distribution with a same covariate x_j. The set of covariates for the both models, Poisson count data and Binomial data, can be different, that's why different parameters are used.

Hurdle models do not distinguish between outliers, structural, or sampling zeros, and assume that all zeros come from a single population, and actually, the formulation of the zero-hurdle models does not tell us the source of zeros. Again, it should be emphasized that these models have two components: a hurdle component for zeros versus non-zeros and a truncated count component for positive counts. Therefore, they model zeros separately from positive counts: these treat the data as a level of presence and absence and analyse the presence data with a count model. In microbial data this is an advantage compared to zero-inflated models, as these models do not differentiate between zeros and fit better, since it is difficult to differentiate zeros into outliers, structural and sampling zeros from conceptual and data generation perspectives.

In ZIP and ZINB models, binomial regression models the probability of a structural zero and/or outlier versus other types of count data, while in ZHP and ZHNB, it models the probability of presence versus absence of a bacterium. Hence, the estimated regression parameters obtained by

ZHP and ZHNB have opposite signs from those obtained by ZIP and ZINB due to the different definition of p_j.

Finally, to carry out these models, applying them to microbiome data, it is necessary to declare an *offset*. The offset will be adjusted as a covariate in the model later to ensure microbiome response is relative abundance instead of count data. This is a very important step to be able to fit models in the study of microbiome data (Xia et al., 2018). The offset argument can also be set only for the count model; so if it takes the link function for truncated model of the zero hurdle model and if the offset is added the equation would look like:

$$\ln(\mu_j) = X_j\beta + \ln(Offset_j) = X_j\beta + offset_j \tag{20}$$

4.2. Model Selection:

As the statistician George Edward Pelham Box, said: *"All models are wrong. Some models are useful"*, it is essential to establish whether the statistical model or models selected, to address a particular issue, are not wrong, since they would lead us to assume useless or incorrect conclusions on important aspects of the hypothesis in question. In most real-world data analysis situations, researchers consider several statistical models that might be appropriate for application. Thus, different criteria need to be addressed in order to choose the model that will give the best result and to which the data will best fit.

Commonly the models: ZIP vs. ZHP and ZINB vs. ZHNB, offer a similar fit to the model and predicted values, but different estimated parameters. The main difference between these models is the interpretation of the parameters, their competence to work with the excess of zeros and in the concept of the zeros that they model. The main issue is to determine which distribution and which model is best suited to deal with this type of data, so each of these should be compared using likelihood ratio test, Akaike's information criterion (AIC), Bayesian information criterion (BIC) and Vuong test (Xia et al., 2018, Xia & Sun, 2017, L. Xu et al., 2015, Hu et al., 2011). Nested models are compared using the likelihood or score test, while non-nested models are evaluated using the AIC, BIC and/or Vuong test. Nested models include ZIP versus ZINB, and ZHP versus ZHNB; and non-nested models include ZIP versus ZHP and ZINB versus ZHNB (Xia et al., 2018).

I. **Akaike's information criterion:** AIC is one of the traditional model-comparison criteria; it can be used for comparing non-nested models (Bozdogan, 1987). The key idea of this criterion is to penalize an excess of adjusted parameters. AIC is used to choose between non-nested mixture models (Xia et al., 2018). It is defined as:

$$AIC(k) = -2\, ln\mathcal{L}\big[\hat{\theta}(k)\big] + 2k \tag{21}$$

Where $\mathcal{L}\big[\hat{\theta}(k)\big]$ is the likelihood function, $\hat{\theta}(k)$ is the maximum likelihood estimate of the parameter vector θ and k is the number of independent parameters estimated within the model, while ln denotes the neperian logarithm (Bozdogan, 1987). The smaller the AIC value, the better the model fit.

II. **Bayesian information criterion:** in the context of procedures based on likelihood, Schwarz in 1978, suggested that the AIC might not be asymptotically justifiable and presented an alternative information criterion based on a Bayesian approach, the BIC, with this criterion penalizing the number of parameters with $ln(n)$, instead of two; expressed as:

$$BIC(k) = -2ln\mathcal{L}\left[\hat{\theta}(k)\right] + (ln\, n)k \qquad (22)$$

Where $\mathcal{L}\left[\hat{\theta}(k)\right]$ is the likelihood function, $\hat{\theta}(k)$ is the maximum likelihood estimate of the parameter vector θ and k is the number of independent parameters estimated within the model, while n is the sample size (Schwarz, 1978) .
As in the previous criterion (AIC), the penalties are used to reduce the effects of overfitting and note that the penalty is more robust for BIC than AIC, regardless of sample size. Because this criterion (BIC) applies a more robust penalty, for the estimation of each additional covariate, it generally selects those models that are more simplified (that is, it chooses the model with the least number of covariates). Because the microbiome data have high heterogeneity (as explained in previous chapters), the criterion that will best select the model will be the BIC (Xia et al., 2018).

III. **Likelihood Ratio test:** this test evaluates the goodness of fit of two nested statistical models based on the relationship of their probabilities, specifically one found by maximization across the parameter space and another found after some restriction is imposed (Kent, 1982).

IV. **Vuong test:** the objective of this test is to compare two non-nested models (f_1 and f_2) that fit the same data through maximum likelihood; the null hypothesis assumed is that the both models fit the data equally well. This test does not require that the models be nested, nor does one of the models need to represent the correct specification (Kent, 1982).

The specific metric of model fit is the Kullback-Leibler Divergence (KLD) from the true model that generated the data (f_t), it is a measure of the distance between two probability distributions and is the basis for other measures of model comparison or selection, such as the AIC (Kent, 1982) . The null hypothesis that dominates this test is:

$$H_0: D_{KL}(f_t||f_1) = D_{KL}(f_t||f_2) \qquad (23)$$

Where f_t and $f_1\ or\ f_2,$ are counting models for non-negative integers. This test is mainly used in models that deal with excess zeros, such as zero-inflated and zero-hurdle models, since it is associated with the test of overdispersion and zero inflation (Xia et al., 2018).

The Vuong test is defined as the average of the logarithmic probability ratio conveniently normalized so that it can be compared with a standard normal:

$$V = \frac{\sqrt{n} * \bar{m}}{S_m} \tag{24}$$

Where n is the sample size, S_m is the standard error of the test statistic, $\bar{m} = \left(\frac{1}{n}\right)\sum_{i=1}^{n} m_i$ and $S_m^2 = \left(\frac{1}{n-1}\right)\sum_{i=1}^{n}(m_i - \bar{m})^2$, and $m_i = \log\left[\frac{f_1(y_i)}{f_2(y_i)}\right]$ (Xia et al., 2018).

If the Vuong test selects two models as valid, the Rootograms and Q-Q Plots of the models will be carried out in order to establish the model that best fits the data:

V. **Rootograms:** The Rootograms are used to display count data regressions, these plots compare observed and expected values graphically by plotting histogram-like rectangles or bars for the observed frequencies and a curve for the fitted frequencies, all on a square root scale. The square roots rather than the untransformed observations are employed to approximately adjust for scale differences across the values or intervals. Otherwise, deviations would only be visible for large observed/expected frequencies. The most popular Rootograms are denominated "*Hanging"*, these align all deviations along the horizontal axis, the bars are drawn from $\sqrt{expected}$ to $\sqrt{expected} - \sqrt{observed}$ so that they are *"hanging"* from the curve representing expected frequencies. The abscissa axis, positioned at 0, will indicate whether the model is over- or under-fitting.

VI. **Q-Q Plots:** in the Quantile-Quantile (or Q-Q) graphs of random versus corresponding theoretical standard quantile residuals, it is checked whether the residuals fit correctly on the line and are between the limits of -2 and 2, thus confirming the hypothesis of normality of the residuals (Dunn and Smyth 1996).

4.3. Models for count data without excess of zeros:

So far, only models dealing with excess zeros have been analysed, but when working with microbiome data, cases with no or very few zeros can also be found, and therefore previous model are not appropriate. Two more methods for working with counts data without excess zeros will be shown below:

Generalized Linear Models (GLM) extend the familiar linear models of regression and ANOVA to include counted data; the Poisson GLM is particularly useful for count data as these tend to be heterogeneous and are always non-negative, and the Negative Binomial GLM has the same advantages as the Poisson distribution with the difference that it allows to model data with overdispersion. A GLM consists of three steps: 1. the distribution of the response variable (where the distribution of Poisson or the Binomial Negative would be specified), 2. the specification of the systematic component in terms of explanatory variables, 3. the link between the mean of the response variable and the systematic part (in this case, the link function will be the logit) (Zuur et al., 2009) . To determine which model best fits the data, it is necessary to use the same criteria explained in the previous section.

In 2014 Love, Huber and Anders proposed a binomial negative model for differential analysis of count data, which is in the *DESeq2* package from RStudio and is based on the methodology developed by Robinson and Smyth in 2010. This package was reviewed as one of the most popular implementations of the variance stabilization technique currently used in RNA Seq-analysis and can be adapted for microbe count data (McMurdie & Holmes, 2014) . This approach allows a valid comparison between OTUs, while substantially improving both power and accuracy in detecting differential abundance (Xia et al., 2018).

These models account for the biological variations in the count data of the high-yield sequencing by means of the mean-variance relationship (Xia et al., 2018):

$$Var(Y_{ij}) = \mu_{ij} * (1 + \mu_{ij}\phi_i) \qquad (25)$$

Where it can be seen that the variance and mean are linked by a local linear regression. Y_{ij} represents the number of readings in sample j that was assigned to OTU_i, then $Y_{ij} \sim NB(\mu_{ij}, \sigma_{ij}^2)$;. ϕ_i is a single proportionality constant that is the same throughout the experiment and can be estimated from the data (McMurdie & Holmes, 2014).

5. Objectives:

The main objective of the present work is to analyse the distribution and behaviour of the "0" values in the large matrix of *omics* data. In order to achieve this general objective, a series of sub-objectives must be raised that will allow us to reach the final goal:

I. Determine when you are working with "excess" zeros.
II. With what maximum percentage of zeros, the models converge.
III. What types of zeros are contained in the bacteria that make up the database.
IV. Select the model that best fits and model the count data with excess zeros and overdispersion.
V. Analyse the behaviour of zeros for categorical covariates, of 2 or more levels.
VI. Carry out the Vuong Test, in different scenarios, to check if you always select the same model as the best.
VII. Determine if the abundance of a given OTU is associated with any environmental or genetic factors.

All these sub-objectives will be carried out and analysed, throughout the Master Thesis.

6. Application to a real database:

All the above models will be tested and applied to two real databases: ***Kraken Counts database*** *(Table 1)* and ***metadataBCLA_RNA*** *(Table 2)*. Both databases come from The Cancer Genome Atlas project (TCGA).

The Cancer Genome Atlas a landmark cancer genomics program, molecularly characterized over 20,000 primary cancers and matched normal samples spanning 33 cancer types. During the following 12 years, the TCGA originated more than 2,5 petabytes epigenomic, genomic transcriptomic, and proteomic source data. All these data are available to any laboratory in the research community. The databases that have been taken from the TCGA to carry out this final master's thesis, are those related to Urothelial Bladder Carcinoma (*BLCA*), (Thorsson et al., 2018).

The first database to be described is ***Kraken Counts database*** *(Table 1),* this database was developed thanks to the *Kraken* tool, this is a popular taxonomic classification tool for metagenomic and microbiome sequencing results.

The database contains a total of 433 muscle-invasive bladder cancer patients included in the TCGA Consortium and 4,263 different bacteria. The bacterial count of a certain type, which each subject contains, is displayed. In addition, the taxa identification number of each bacterium and the sum of the counts that each one has in the 433 patients is shown. As it can be seen, there is a very large difference between some bacteria and others, while some present more than 5 million counts across the 433 patients, others present a very low total of 10 counts or even a single count across the subjects. ***Kraken Count database*** contains both bacteria that were found in bladder tumours and bacteria that are part of our usual microbiome (of the urinary tract, intestinal tract, skin, ...).

Table 1. *Summary table of the Kraken Counts database.*

taxID	Id_1	Id_2	Id_3	...	Id_{433}	*Specie*	*Total*
96344	0	16	0	⋮	0	*Cupriavidus oxalaticus*	9405657
1491	0	41092	2	⋮	0	*Clostridium botulinum*	2621385
2184519	1	19060	0	⋮	33	*Hydrogenophaga sp. NH-16*	1496395
1358	0	2	0	⋮	0	*Lactococcus lactis*	333344
1396	0	3012	27	⋮	1	*Bacillus cereus*	324567
1747	18	531	28	⋮	31	*Cutibacterium acnes*	128398
1286	4	0	33	⋮	374	*Staphylococcus simulans*	47105
339	0	53	1	⋮	0	*Xanthomonas campestris*	33067
573	81	81	50	⋮	53	*Klebsiella pneumoniae*	31519
1275	0	4	1	⋮	0	*Kocuria rosea*	22691
550	82	71	14	⋮	42	*Enterobacter cloacae*	21275
485	80	55	21	⋮	47	*Neisseria gonorrhoeae*	19264
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1003110	0	0	0	⋮	0	*Verrucosispora maris*	0

MetadataBCLA_RNA contains a total of 119 different variables, of the same 433 muscle-invasive bladder cancer patients included in the TCGA Consortium and sampled to form *the Kraken Count database*. This database is the result of merging 3 different databases: it contains 39 different gene expressions, 75 variables related to the immune system response and the rest are either tumour characteristics (such as stage, sample type, tissue or organ of origin, site of resection or biopsy, ...) or subject-related characteristics (such as gender, race, weight, BMI, number of cigarettes smoked per day, etc.) All these variables are related to bladder cancer. For example, within the 39 gene expression variable, 10 are called Toll-Like Receptors (TLRs) and they play key roles against cancer (Ohadian Moghadam & Nowroozi, 2019). Within the variables, classified as immune system response, several are related to the T Cell Receptor (TCR) and B Cell Receptor (BCR), the main types of lymphocytes, T cells and B cells, surface receptors that recognize antigens, are activated to initiate an immune reaction in response to the specific binding of their receptors to antigens such as tumours and viruses (Gagnaire et al., 2017).

Therefore, in this project it will be analysed whether there is a certain relationship between the bacteria and some of the variables in the ***MetadataBCLA_RNA*** database, while studying and modelling the behaviour of the 0 values. One of the problems with the ***MetadataBCLA_RNA*** database was the large number of missing data contained in some variables, so the first step was to calculate the percentage of missing in each variable *(Table 2, Figure 5)*. As a result of *Figure 4*, it was found that depending on the origin of the variable (if it is a genetic expression, an immune system response, …) it presented the same percentage of lost values but not in the same subjects:

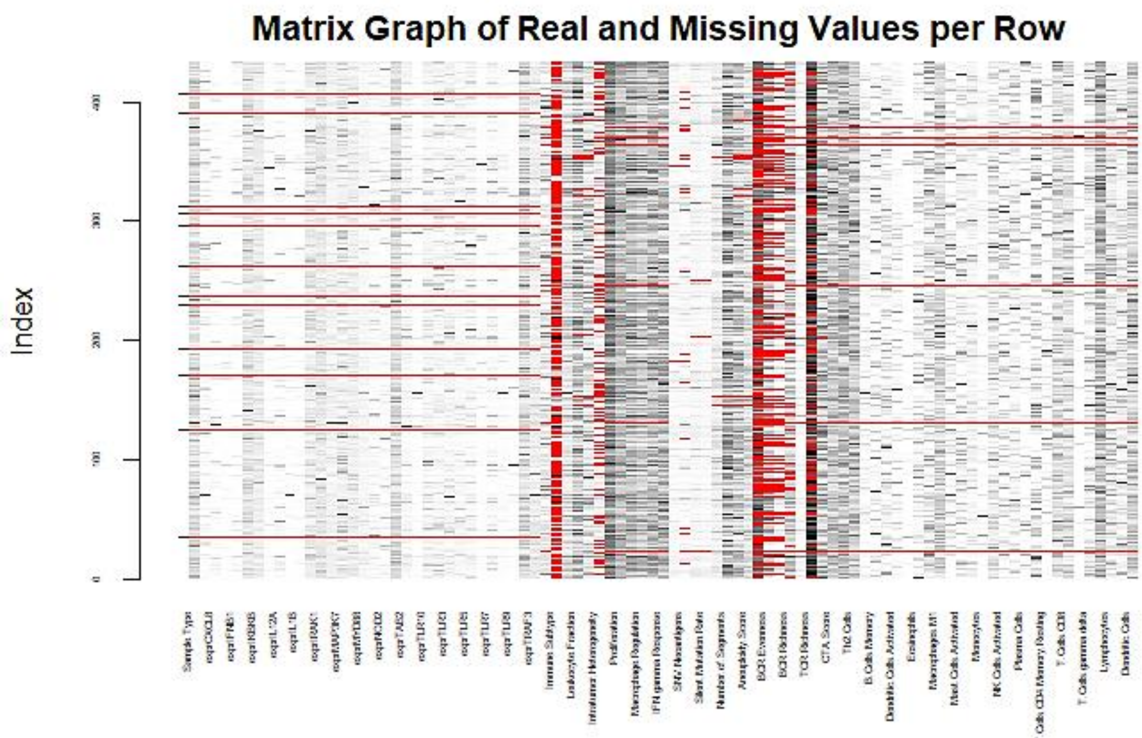


Figure 5. *Matrix Graph of missing and observed values, by rows. In the X axis are reflected the 433 subjects of the database and in the Y axis some of the variables, the red lines show the missing values that each one of the subjects has in the variables. Image created with the VIM package of the RStudio software.*

In the first 18 variables, which are different variables of gene expression, it can be seen that it is always the same subjects who present NA (the red lines), but as soon as one passes to variables related to the immune system or environmental factor, the subjects who presented NA in the first variables no longer present it in the rest of the variables. Therefore, it is not always the same subjects who present NA throughout the entire database.

Table 2. *Summary table listing the name of the variable and the percentage of missing values it has in total. The complete table can be found in Annexes, page 9.*

Variable	***Percentage of missing values (%NA)***	***Median***	***Var***	***IQR***
exprCHUK	4.389	-0.389	1.113	1.281
exprCXCL8	4.389	-0.415	1.329	0.537
⋮	⋮	⋮	⋮	⋮
TCGA.Subtype	65.589	BLCA.2	-	-
Stromal.Fraction	2.541	0.4	0.051	0.371
Proliferation	3.234	0.559	0.274	0.595
⋮	⋮	⋮	⋮	⋮
BCR.Evenness	46.189	0.884	0.051	0.125
BCR.Richness	46.028	9	1555.325	31
⋮	⋮	⋮	⋮	⋮
BMI	14.319	24.919	41.342	6.643

An attempt was made to carry out a missing pattern to analyse whether these were MAR, MNAR or MCAR type, but due to the large number of variables this was very complicated or almost impossible to analyse. Finally, we chose to select those variables that had the lowest percentage of missing values and that allowed us to study (from all approaches) the behaviour of the 0-values, as it is a methodological project, all possible cases will be studied: quantitative variables with NA and without NA, categorical variables with two or more levels, with NA and without NA *(Table 3)*.

The imputation of the variable Body Mass Index (BMI) was carried out, due to the good results obtained when using the *predictive mean matching* technique (*Annexes*, *page 12*). For each missing entry, the method forms a small set of donor candidates from all the complete cases that have predicted values closest to the predicted value for the missing entry. A donor is randomly drawn from the candidates and the observed donor value is taken to replace the missing value. The distribution of the missing cell is assumed to be the same as the observed data for the candidate donors (Schenker & Taylor, 1996). Once the values have been allocated, they are categorized into four different levels, according to the subject's kilograms: *Underweight* (BMI is less than 18.5) - *Normal weight* (BMI is 18.5 to 24.9) - *Overweight* (BMI is 25 to 29.9) – *Obese* (BMI is 30 or more).

Table 3. *Summary table of selected variables. The following table shows the variables that have been selected to carry out the models.*

Variable	*Variable type*	*Percentage of missing values (%NA)*	*Median*	*Var*	*IQR*
exprTLR7	Numerical	4.389	-0.336	5.719	0.659
exprMAP3K7	Numerical	4.389	-0.458	1.605	1.352
exprNOD1	Numerical	4.389	0.009	1.885	0.866
Leukocyte.Fraction	Numerical	0	0.202	0.027	0.233
			Category levels		
Immune. Subtype	Categorical	3.233	C1, C2, C3, C4, C6		
Tumor. Stage	Categorical	0.462	Stage i, Stage ii, Stage iii, Stage iv		
Gender	Categorical	0	Male, Female		
BMI	Categorical	0	Under, Normal, Overweight, Obese		

The workflow that will be carried out to be able to determine all the established hypotheses and study the behaviour of the 0 values, will be the following *(Figure 6):*

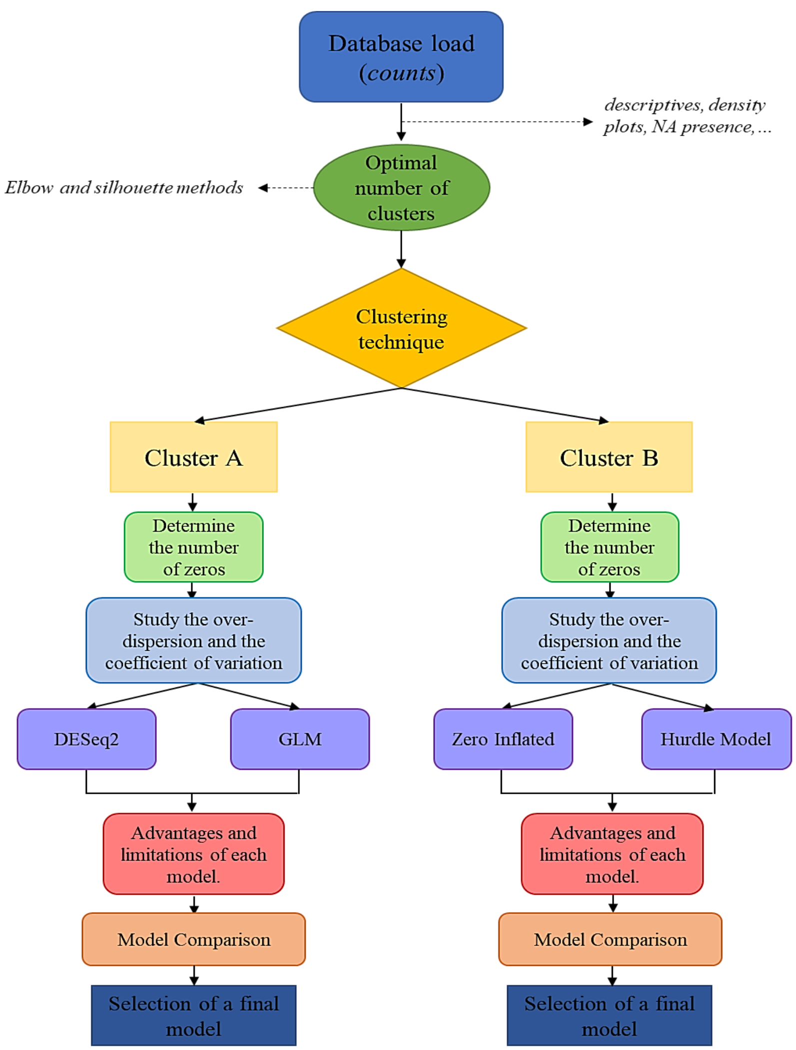


Figure 6. *Workflow. These are the steps that will be carried out to meet the main objective of the work. Image created in Biorender.com*

As mentioned at the beginning of section 6 *(6. Application to a real database)*, ***Kraken Counts Dataset*** contains 4,263 bacteria, and each of these behaves differently in different subjects. If density plots are made of these bacteria *(Figure 6),* it can be seen that they are different from each other, and that each one of them has a different number of zeros. While there are bacteria that have a high percentage of zeros and have a high total positive count (as is the case of *Streptococcus mitis*, which has 76.7% zeros and 23.3% positive values, corresponding to a total of 417 positive values); there are others that are just the opposite, a high percentage of zeros and a low total positive count (as is the case of *Muricauda ruestringeis* which has 99.9% zeros and 0.1% positive values, corresponding to a total of 3 positive counts).

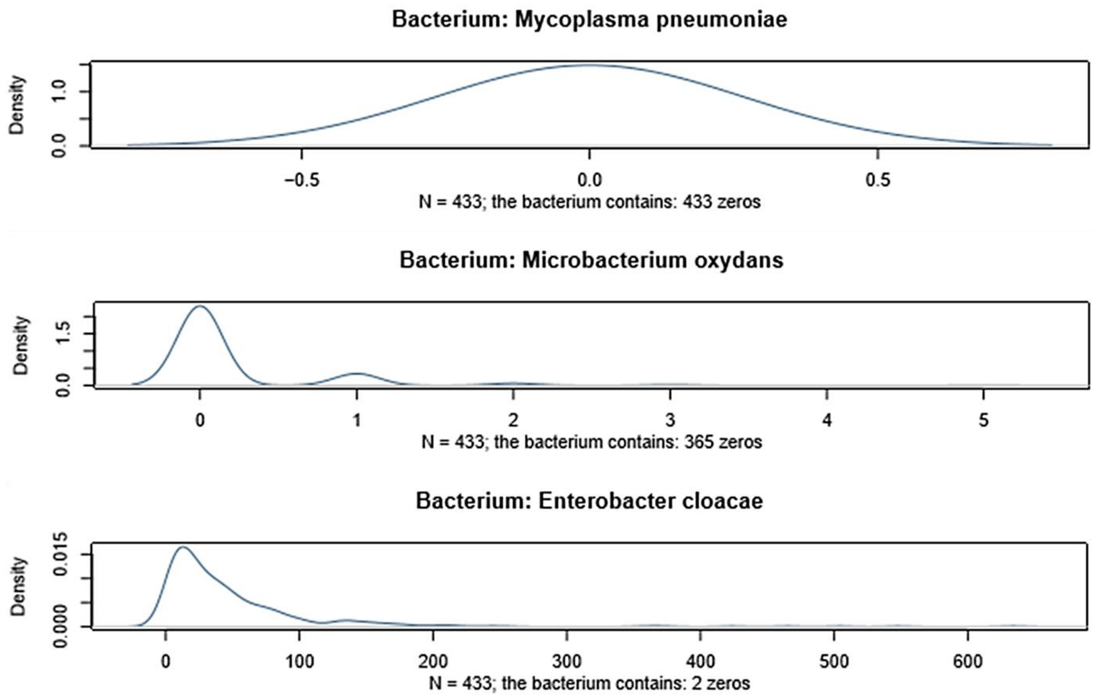


Figure 7. *Density graph of some of the bacteria in Kraken's database. These three bacteria are taken at random from the document showing the density graphs of all the bacteria; the title indicates the name of the bacteria and below the graph, the number of 0's it contains. Image created with the ggplot2 package of the RStudio software.*

As it is specified in *Section 3.1. Models for count data with excess of zeros* and *3.2. Models for count data without excess of zeros*, depending on the number of zeros in the variable, one method or another is applied. The main question is from what percentage of zeros it can be considered an excess of zeros. The authors working in this field are always referring to an excess of zeros, but they do not specify which is the cut-off point. So, the first key point is to determine, which bacteria have an excess of zeros and which do not.

Instead of establishing a cut-off point, which was not based on any scientific evidence, a clustering approach was proposed so that those bacteria with more similar counts would be grouped together. The proposed clustering method is *k-means* algorithm which allows to group, since it allows to group objects into k groups based on their characteristics. The grouping is done by minimizing the sum of distances between each object and the centroid of its group or cluster. Before carrying out the algorithm, two different tools will be applied to calculate the optimal number of clusters into which the database would be divided:

- <u>The Elbow Method</u> *(Figure 8)* is one of the most popular methods for determining this optimum value from a range of *k*-values. This approach plots the value of the cost function produced by different *k*-values, this function decreases as *k* increases, represented as the Total Sum of Squares:

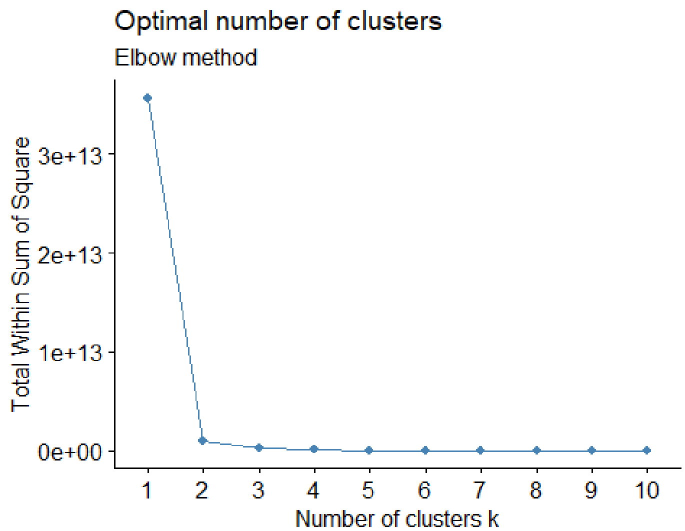


Figure 8*. Elbow method. In this graph on the Y axis the Total within sum of square and on the X axis the number of k clusters. Graph created with the "factoextra" and "NbClust" packages of the RStudio software.*

- <u>The Silhouette Method</u> *(Figure 9)* measures of how similar an object is to its own group compared to other groups. Silhouette scores are in the range of [0, 1], a value of 1 indicates that the sample is far from its neighboring group and very close to the assigned group and, a value of 0 means that the distance between groups is equal.

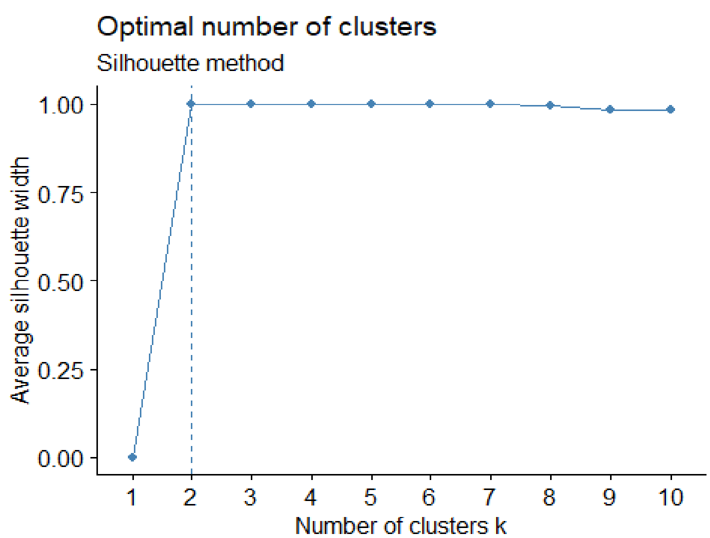


Figure 9. *Silhouette method. In this graph on the Y axis the average silhouette width is indicated and on the X axis the number of k clusters. Graph created with the "factoextra" and "NbClust" packages of the RStudio software.*

As it can be seen the optimal number of clusters is 2 in the *Elbow Method*, from that point the increment is almost null, and in the *Silhouette Method* the maximum score is reached with two clusters, which would indicate that the sample is well grouped and assigned to a highly appropriate group.

Once the optimal number of clusters has been checked, the *k-means* algorithm is carried out *(Figure 10)*, where two different groups can be observed. The blue cluster contains 4,257 bacteria and the pink cluster contains only 6. If the number of zeros in each of the bacteria in each group is analyzed, it can be seen that the bacteria with and without an excess of zeros, have been automatically brought together. This step is crucial to identify which methodology is used in each case.

Figure 10. *K-Means Clustering. The blue cluster contains all the variables with excess of zeros and the pink cluster, those that do not have this characteristic. Graphic created with the "NbClust" package of the RStudio software.*

For example, some of the bacteria in the pink cluster do not even have a single zero, e.g. *Cutibacterium acnes* or *Klebsiella pneumoniae*, and others have 12.24% of zeros, as *Staphylococcus simulans* (which contains the highest percentage of zeros in the pink cluster). While in the blue cluster more than 1,000 bacteria only contain zeros, another 2,017 bacteria have 99% ~ 98% of zeros. Most bacteria are characterized by more than 70% of zeros.

From now on, different methodologies will be applied to each group. For the blue cluster we will use the appropriate methods for dealing with excess zeros (Zero Inflated Model and Zero Hurdle Model). For the pink cluster it will be use methods that do not work with this characteristic (*DESeq2* and General Linear Models).

6.1. Cluster with excess zeros:

In this section the group with excess zeros will be analysed in order to determine if the abundance of any bacteria is related to any environmental or genetic factor; therefore, as a dependent variable, the different bacteria in the ***Kraken Counts database*** will be taken, and as independent variables, the variables selected in ***MetadataBCLA_RNA*** *(Table 3)*. Both Zero-Inflated and Zero-Hurdle models are implemented in *RStudio*. In this work it will be used the *"pscl"* package (Zeileis at al., 2008). The total count read to create the offset is taken. In *Frame 1* it shows the R-code used to fit these models (all the databases are sorted by the subject ID):

```
#Create the offset with the total reads of bacteria from the Kraken
count dataset:
clst2$Offset <- log(totalReadsKraken$nonHumanReads)
# The first part of the formula (Kraken) displays the bacteria and the
second (metadata) the covariates:
fm < - formula (Kraken[, i] ~  metadataBCLA_RNA[,j] + offset(Offset))
```

Frame 1

TotalReadsKraken contains the total read counts and belongs to the same individuals found in ***Kraken Count Database*** and in ***MetadataBCLA_RNA***. The zero-inflated models will come by, *Frame 2*:

```
ZIP < - zeroinfl(formula = fm, dist = "poisson", link = "logit", data
 = metadataBCLA_RNA)

ZINB <- zeroinfl(formula = fm, dist = "negbin", link = "logit", data
 = metadataBCLA_RNA)
```

Frame 2

The *dist* option specifies the distribution for the count data. The *link = logit* option specifies the logistic link. The same formula as the previous ZIP and ZINB models are used to adjust the ZHP and ZHNB models using the *hurdle()* function, *Frame 3*:

```
ZHP <- hurdle(formula = fm, dist= "poisson", data = metadataBCLA_RNA)

ZHNB <- hurdle(formula = fm, dist= "negbin", data = metadataBCLA_RNA)
```

Frame 3

Figure 11, shows the distribution of zeros of the dataset. Scenarios are shown in which one will study how zeros behave, which distribution/model fits best and from which percentage of zeros and counts one starts to have reliable results. In each of the different scenarios, a bacterium (that meets the conditions described) will be used to carry out the examples:

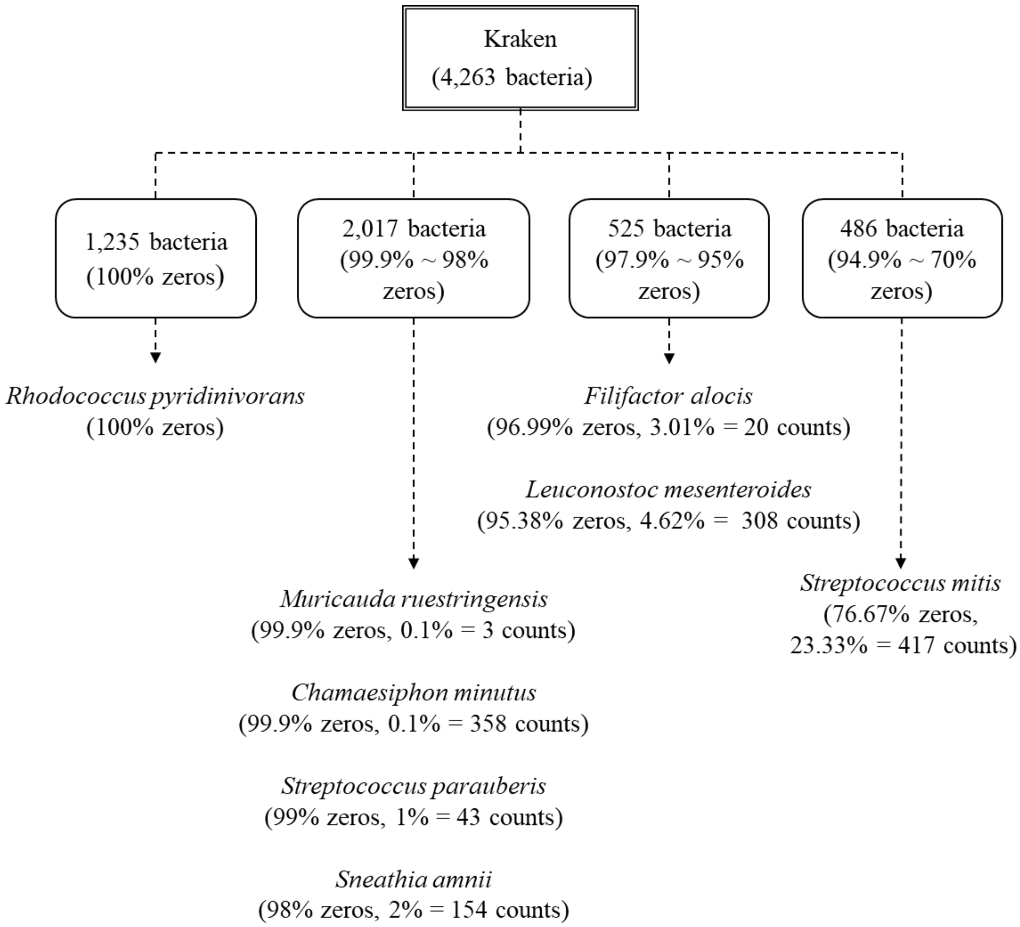


Figure 11. *Kraken dataset zero distribution. This diagram shows the percentage of zeros that this database has, and the number of bacteria that are characterized by it. Graph created in Biorender.com.*

I. Bacteria with 100% zeros

In ***Kraken Count dataset***, there are a total of 1,253 bacteria containing only zeros, *Rhodococcus pyridinivoras* is one of the bacteria that meets this condition. When some bacterium with 100% zeros is introduced into the models, as a dependent variable, the model has no sense. Therefore, all those bacteria with 100% zeros are removed from the database.

II. Bacteria with 99.9% ~ 98% zeros

The ***Kraken Counts dataset***, contains 2,017 bacteria with 98% to 99.9% zeros, therefore, in this item, several cases will be studied (in all models it will be used, as an independent variable, the leukocyte fraction):

⇒ ***Case 1:*** bacteria with 432 zeros and a total of 3 positive counts.

Muricauda ruestringensis contains 432 zeros, and one subject has 3 counts of this bacterium:
$\mathbb{E}_{M.Ruestringensis} = 0.007$, $\mathbb{V}ar_{M.Ruestringensis} = 0.021$, $\mathbb{CV}_{M.Ruestringesis} = 2081.63\%$

The variation has been found to be larger than the expected value of the data, indicating a clear over-dispersion. In addition, such a high coefficient of variation (CV) is indicating that the data are highly variable, i.e., they are "heterogeneous" and highly dispersed. If a Zero Inflated Model is used, these models do not yet converge due to the large number of zeros in the data; if a Zero Hurdle Model is used, happens again as in the previous scenario: the model still does not converge. Therefore, with 99.9% of zeros and such low positive counts, no valid results are obtained either. Bacteria with these characteristics are also eliminated from the database.

⇒ ***Case 2:*** bacteria with 432 zeros and a total of 358 positive counts.

The next scenario to be contemplated is a bacterium with the same percentage of zeros as the previous one but with a much higher number of positive counts, as is the case of *Chamaesiphon minutus* with 432 zeros and 358 counts:

$$\mathbb{E}_{C.Minutus} = 0.827, \qquad \mathbb{V}ar_{C.Minutus} = 295.991, \qquad \mathbb{CV}_{C.Minutus} = 2080.87\%$$

In this case the variance is much higher than the expected value of the data, indicating a high over-dispersion of the data. This CV is very similar to the coefficient of the previous case, so it would be interpreted in exactly the same way: the data are highly dispersed, they are "heterogeneous".

This time the zero inflated models do converge but give the same "error" as the zero hurdle model: they still estimate missing values. Therefore, although both models converge, they do not give valid results, so regardless of the number of positive counts these bacteria have, those with 99.9% of zeros are eliminated from the database.

⇒ ***Case 3:*** bacteria with 429 zeros and a total of 43 positive counts.

It continues to choose a bacterium with a lower percentage of zeros, *Streptococcus parauberis*, with 429 zeros (99%) and a total of 43 counts:

$$\mathbb{E}_{S.Parauberis} = 0.099, \qquad \mathbb{V}ar_{S.Parauberis} = 2.784, \quad \mathbb{CV}_{S.Parauberis} = 1680.20\%$$

As in the previous contexts, the data still show overdispersion, with the difference that the CV is lower than that of the previous bacterium, although the interpretation is the same: the data are still highly dispersed, they are clearly "heterogeneous".
Again, the models are carried out, with the difference that for the first-time results are obtained in the count model coefficients section, so that from this percentage of zeros valid results are obtained for the models:

- Zero Inflated Poisson model: *(Output 1 – Annexes, page 14)*

- Zero Inflated Negative Binomial: *(Output 2 – Annexes page 14)*

- Zero Hurdle Poisson model: *(Output 3 – Annexes page 15)*

- Zero Hurdle Negative Binomial model: *(Output 4 – Annexes page 15)*

By default, *Outputs 1 to 4* show estimated coefficients, standard errors, values for the Wald test and associated *p-values*, but no confidence intervals. As a main observation, the zero component has not only the estimated parameters different in magnitude, but also their signs reversed. The difference sings between Zero Inflated Model and Zero Hurdle Model is due to the *hurdle()* function modelling the probability of a non-zero count, instead of the probability of a zero count.
The model selection criteria are carried out *(Table 4),* before establishing which model is best (applying the Vuong test), define which distribution best fits the data, if a Poisson or a Negative Binomial, using the AIC and BIC criteria along with the Likelihood Ratio Test (the lower value will indicate which distribution the data best fit):

Table 4. *Model comparison based on AIC, BIC and likelihood ratio test. In this table, the results for the AIC and BIC criteria and the result of the likelihood ratio test (LHRT) are shown for each of the above models, with the Streptococcus parauberis bacteria and the covariate Leukocyte. Fraction.*

Models	*AIC*	*BIC*	*LHRT*
ZIP	96.874	113.157	ZINB
ZINB	76.747	97.101	
ZHP	96.997	113.280	ZHNB
ZHNB	76.484	96.838	

As has been proven, the distribution that best fits the data is the Negative Binomial. A very important point to consider, is the *Output 2 and Output 4* of each of these models and how it would be interpreted. The correct parameter interpretation should be based on the model definitions. As defined, the logistic component in the Zero Inflated Models corresponds to inferences about the outliers and structural zero groups; in the contrast, the logistic component in the Zero Hurdle Model correspond to inferences about zeros, in general *(Section 4.1. Models for count data with excess of zeros).* The parameters from the count component can be elaborated this way: for the log-linear component in the Zero Inflated model, parameters are interpreted with respect to the non-outlier and non-structural zero group, whereas for the Zero Hurdle model they are interpreted with respect to the non-zero group. In ZINB, it find that the log odds of being in the non-structural zero group (those who can have counts *S. Parauberis)* decrease with 7.142; on the level of the odds this means a decrease by $\mu_i = \exp(-7.142) = 0.00079$. For the count component of the ZHNB model (that is, parameter estimates conditioned on a subject having a positive count of *S. Parauberis* the probability of having at least one positive count of *S. Parauberis is,* $\mu_i =$

$\exp(-5.3305) = 0.0048$. Following the previous result *(Table 4)*, the two finalist models are evaluated with the Vuong test *(Outcome 5):*

```
Model 1
 Class: zeroinfl
 Call: zeroinfl(formula = fm, data = metadataBCLA_RNA, dist = "negbin", ...

Model 2
 Class: hurdle
 Call: hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")

Variance test
  H0: Model 1 and Model 2 are indistinguishable
  H1: Model 1 and Model 2 are distinguishable
    w2 = 0.001,   p = 1

Non-nested likelihood ratio test
  H0: Model fits are equal for the focal population
  H1A: Model 1 fits better than Model 2
    z = -0.470,   p = 0.681
  H1B: Model 2 fits better than Model 1
    z = -0.470,   p = 0.319
```

Outcome 5

The null hypothesis testing by Vuong test is that both models are indistinguishable. Based on the *p-value*, it is concluded that there is not enough evidence to reject it. Because the Vuong test does not determine which model is better, the residuals and Rootogram of each of the models are analysed *(Figure 12):*

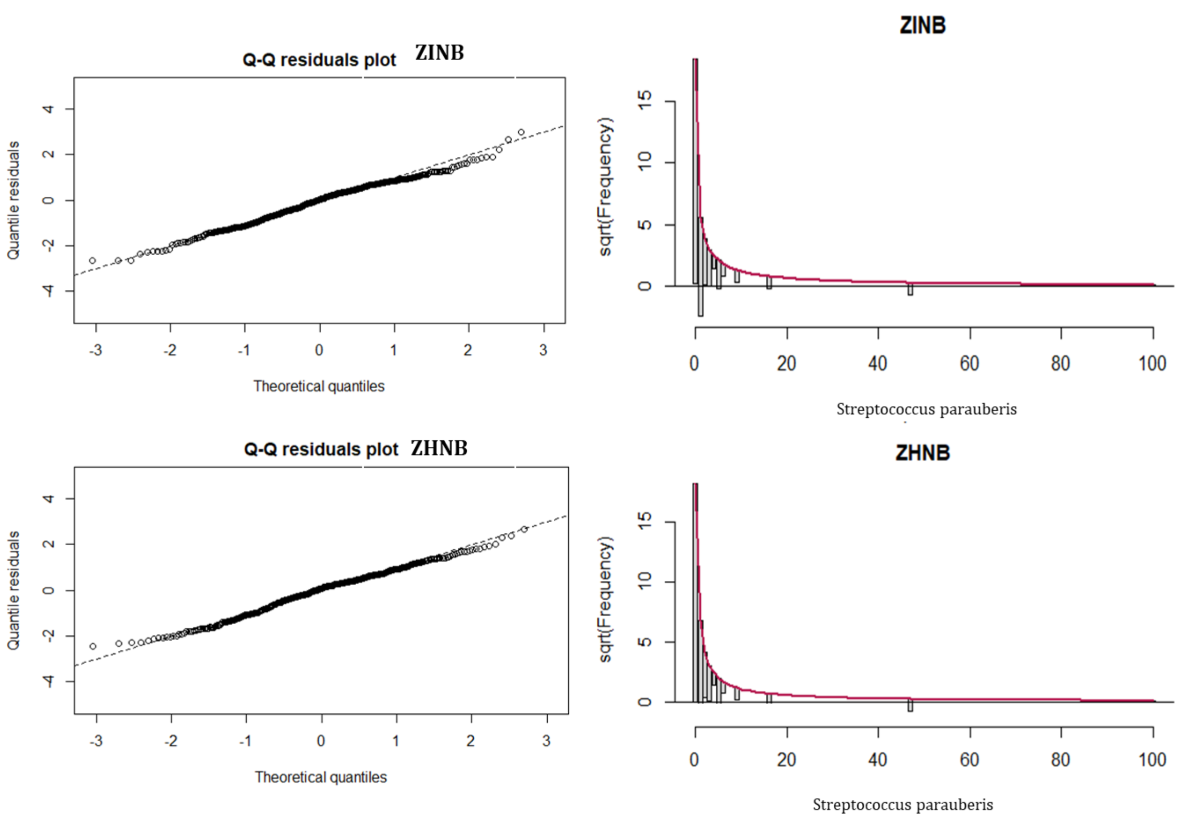


Figure 12. *Q-Q Residuals Plot and Rootogram Zero Hurdle Negative Binomial model. In the graph on the left are the conditional Pearson residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

As explained in section *3.2. Model Selection*, in the Rootograms the line at 0 allows us to easily visualize where the model is over- or under-fitting, so at 0 it fits perfectly by design. In the first counts of the Zero Inflated model we see a mismatch (below the line), which does not occur in the Zero Hurdle model.

Both Q-Q Plots are checked to see if they fit the line correctly and are between -2 and 2, confirming the hypothesis of normality of the residuals but at the right end of the Zero Inflated model it can be seen how the residues "move away" from the line. Therefore, through these graphs, it was concluded that the Zero Hurdle Model fits the data better. In addition, there is scientific evidence that the barrier model is better in terms of microbiome data (since it does not differentiate between zero types and has greater convergence power than zero inflated model), which is just what was determined with the previous Outputs and plots. If all the models are performed again but using another type of covariate such as the expression of a certain gene, the gender of the subject, etc., the Vuong test provides the same result.

If the independent variable is a two level categorical variable, such as gender, the models does return a result. However, if the covariate has more than two levels, such as *Immune. Subtype*, *Tumor_stage* and *BMI*, there is more than one category that includes only zeros and the standard error cannot be estimated *(Table 5)*:

Table 5. *Cross table of S. Parauberis with Immune Subtype. This table shows each of the categories of this covariate and the number of zeros and counts that each of them containss*

S.Parauberis	***C1***	***C2***	***C3***	***C4***	***C6***	***NA***
Zero	184	169	22	37	3	14
Positive count	0	3	0	1	0	0

No significant result is obtained for the Zero Hurdle Model coefficients part and for the count model coefficients section, since the number of counts per category is so low of positive values and more than one contains only zeros, the error cannot be estimated, neither the Z-value or the *p-value.*

⇒ ***Case 4:*** bacterium with 423 zeros and a total of 154 positive counts.

Sneathia amnii has a 98% zeros, and 2% positive counts (154 total positive values), in order to check if the Vuong test gives different results in relation to which model to work with and if the categorical variables with more than two levels give results in the models:

$$\mathbb{E}_{S.Amnii} = 0.356, \qquad \mathbb{V}ar_{S.Amnii} = 46.609, \qquad \mathbb{CV}_{S.Amnii} = 1919.57\%$$

Again, the models are made with Leukocyte Fraction as a covariate, and the same conclusions are reached as for the previous bacterium, which contained a higher percentage of zeros *(Table 6)*:

Table 6. *Model comparison based on AIC, BIC, likelihood ratio test and Vuong test. In this table, the results for the AIC and BIC criteria and the result of the likelihood ratio test are shown for each of the above model, with Sneathia amnii bacteria and the covariate Leukocyte. Fraction.*

Models	*AIC*	*BIC*	*LHRT*	*Vuong test*
ZIP	900.104	916.387	ZINB	ZINB/ ZHNB
ZINB	136.834	157.187		
ZHP	901.028	917.311	ZHNB	
ZHNB	132.400	152.754		

The Vuong test returns the same result as before, both models fit equally well, but always using a binomial negative distribution as the data are over-dispersed. The zero-hurdle model is chosen because it does not distinguish between zeros and is recommended for working with microbial data (besides being the model with the lowest AIC and BIC). The model is presented in *Output 6*:

```
Call:
hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")
Count model coefficients (truncated negbin with log link):
                        Estimate     Std.Error     z value     Pr(>|z|)
(Intercept)              -41.01        57.10       -0.718      0.47261
Leukocyte.Fraction        61.86        20.44        3.027      0.00247 **
Log(theta)               -10.82        56.72       -0.191      0.84871

Zero hurdle model coefficients (binomial with logit link):
                        Estimate     Std.Error     z value     Pr(>|z|)
(Intercept)              -4.4209       0.6367      -6.944      3.81e-12 ***
Leukocyte.Fraction        2.1169       1.7927       1.181      0.238
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 0
Number of iterations in BFGS optimization: 132
Log-likelihood: -61.2 on 5 Df
```

Output 6

Interpretation: *(Zero-hurdle model)* For those subjects with minimal values in the *Leukocyte Fraction*, the probability of having a positive count of *Sneathia amnii* is 0.012:

$$P\left(S.\widehat{Amnu} \geq 1 | minLeukcyte\ Fraction\right) = \frac{e^{-4.4209}}{(1+e^{-4.4209})} = 0.0118 \sim 0.012$$

(Positive count model) For each 0.1 increase in the *Leukocyte Fraction*, there is a ($e^{61.86 * 0.1} = 485.9$) 485.9 increase in the *Sneathia amnii* rate. The theta count confirms that the data does present over-dispersion. So the link function (taking into account *Equations 10 and 11*) would look like this:

- For the logistic regression:

$$\left(\frac{p_j}{1-p_j}\right) = e^{-4.4209} = 0.012$$

➢ For the truncated model, which will be given with an increase in μ_j:

$$(\mu_j) = e^{61.86} * Leukocyte.Fraction = (7.336 * 10^{26}) * Leukocyte.Fraction$$

The model residuals and Rootogram are shown below *(Figure 13)*:

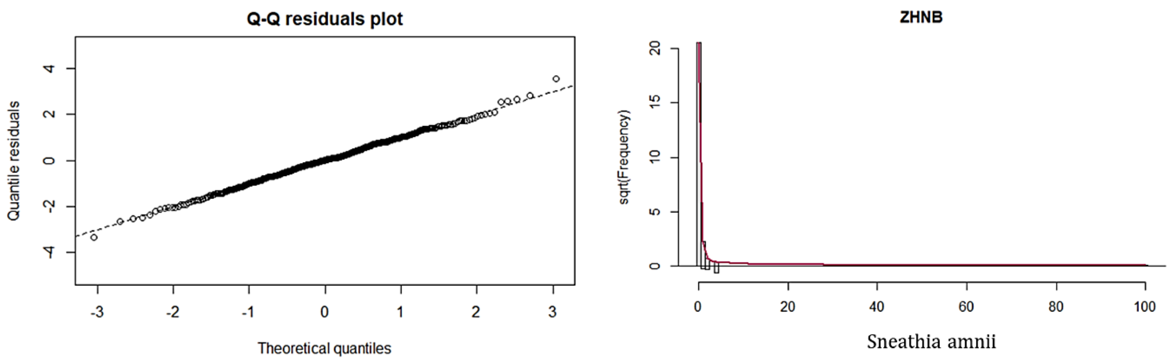


Figure 13. *Q-Q Residuals Plot and Rootogram zero hurdle Negative Binomial model. In the graph on the left are the conditional Pearson residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

In the Q-Q residues graph, it is checked that these fit correctly to the line and are between -2 and 2, confirming the hypothesis of normality of the residues. In the Rootogram it is observed that the expected values (the red line) fit correctly to the observed values, so it could be concluded that the model fits correctly the data.

From this last block it can be concluded that from 99% of zeros, the Zero Inflated model and the Zero Hurdle Model do converge, so that percentage is assumed as a cut-off point through which statistically significant results begin to be obtained.

III. Bacteria with 97.9% ~ 95% zeros

To check if, with a lower percentage of zeros, categorical covariates with more than two levels can be correctly estimated in the model and if the Vuong test still gives the same result as in the previous cases, different cases will be shown again:

⇒ ***Case 1***: bacterium with 420 zeros and a total of 20 positive counts:

The first bacteria to be used is *Filifactor alocis*, with a total of 420 zeros (96.99% zeros) and 20 total counts:

$$\mathbb{E}_{F.Alocis} = 0.0462, \qquad \mathbb{V}ar_{F.Alocis} = 0.146, \qquad \mathbb{CV}_{F.Alocis} = 827.27\%$$

In this case the models will be carried out using two-level categorical variables, i.e. *Filifactor alocis* $\sim \beta_0 + \beta_1 * Gender + offset$; and categorical variables with 5 levels and missing values: *Filifactor alocis* $\sim \beta_0 + \beta_1 * Immune.Subtype + offset$. Before carrying out the models, the zeros and positive counts for each covariate are determined *(Table 7 and 8)*:

Table 7. *F. Alocis bacterium with Gender. This table shows each of the categories of this covariate and the number of zeros and counts that each of them contains.*

F.Alocis	***Male***	***Female***
Zero	307	115
Positive count	8	3

Since all categories of the covariate "gender" have positive counts, the results of all models can be estimated and compared *(Table 8)*:

Table 8. *Model comparison based on AIC, BIC, likelihood ratio test and Vuong test. In this table, the results for the AIC and BIC criteria and the result of the likelihood ratio test are shown for each of the above models, with Filifactor alocis bacteria and the covariate "Gender".*

Models	*AIC*	*BIC*	*LHRT*	*Vuong test*
ZIP	149.985	166.268	ZINB	ZINB/ ZHNB
ZINB	149.972	170.325		
ZHP	148.261	164.544	ZHNB	
ZHNB	145.128	165.481		

Again, the same conclusions as Block II are obtained: the distribution that best fits the data is the Negative Binomial, due to the over-dispersion that the variable shows; the data is adjusted to a Zero Hurdle Model *(Output 7):*

```
Call:
hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")
Count model coefficients (truncated negbin with log link):
                      Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)            -35.50        285.48        -0.124        0.901
Male                    12.19        201.73         0.060        0.952
Log(theta)             -10.19        201.69        -0.051        0.960

Zero hurdle model coefficients (binomial with logit link):
                      Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)           -3.6463        0.5848        -6.235        4.52e-10 ***
Male                   0.2286        0.6673         0.343        0.732
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 0
Number of iterations in BFGS optimization: 88
Log-likelihood: -67.56 on 5 Df
```

Output 7

In the *Zero-hurdle model* coefficients, the probability (for a woman) of having at least 1, *Filifactor alocis* is 0.026:

$$P\left(F.\widehat{Alocis \geq 1}|Women\right) = \frac{e^{-3.6463}}{(1+e^{-3.6463})} = 0.0258 \sim 0.026$$

If, for example, we wanted to calculate the probability of having a positive count for a man on *Filifactor alocis,* it would add to the value of e the man's intercept plus the woman's.

So the link function (taking into account *Equations 11*) would look like this:

- For the truncated model, there would be an increase in μ_j of:

$$(\mu_j) = e^{-3.6463}$$

Because the other coefficients are not statistically significant, none of these can be concluded. A low theta value confirms that the data is over dispersed.

As in previous scenarios, the model residuals and Rootogram are studied *(Figure 14):*

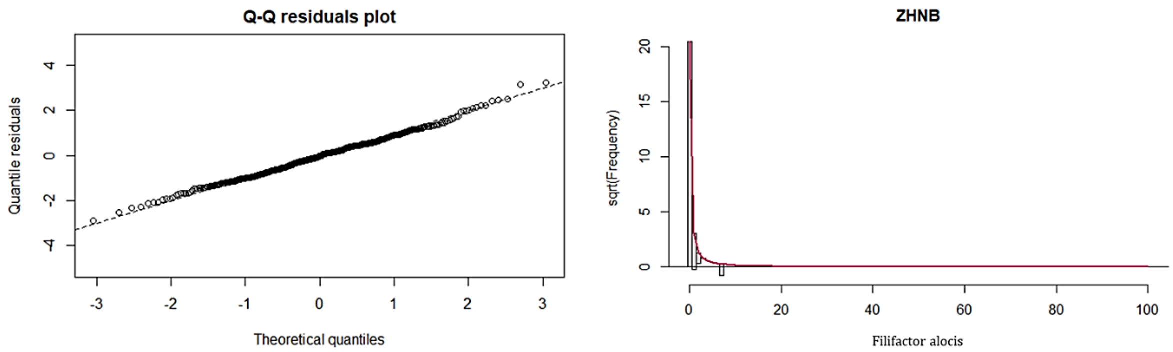


Figure 14. *Q-Q Residuals Plot and Rootogram zero hurdle Negative Binomial model. In the graph on the left are the residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

In the Q-Q residuals plot, it is checked that these fit correctly to the line and are between -2 and 2, confirming the hypothesis of normality of the residues. In the Rootogram it is observed that the expected values (the red line) fit well to the observed values, so it could be concluded that the model fits correctly to the data.

Applying this same model to the Immune subtype variable, it is verified that, presenting 0 positive counts in certain categories and such low positive counts in others *(Table 9)*, the model returns missing values in the count model coefficients part.

Table 9. *Cross table of F. Alocis bacterium with Immune Subtype. This table shows each of the categories of this covariate and the number of zeros and counts that each of them contains.*

F.Alocis	*C1*	*C2*	*C3*	*C4*	*C6*	*NA*
Zero	179	169	21	36	3	12
Positive count	5	3	1	2	0	2

⇒ ***Case 2:*** bacterium with 413 zeros and a total of 308 positive counts.

Leuconostoc mesenteroides, which has fewer zeros and a higher total count than the previous bacterium:

$$\mathbb{E}_{L.Mesenteroides} = 0.711, \qquad \mathbb{V}ar_{L.Mesenteroides} = 123.812, \quad \mathbb{CV}_{L.Mesenteroides} = 1564.27\%$$

Again, it will be analysed which model fits better, this time using a covariate containing NA, to confirm that with this percentage of zeros and with a covariate with missing values, the Vuong test still returns the same result; the selected covariate is the expression of TLR7 with 4.39% of missing values *(Table 10)*:

Table 10. *Model comparison based on AIC, BIC, likelihood ratio test and Vuong test.*

Models	*AIC*	*BIC*	*LHRT*	*Vuong test*
ZIP	1864.285	1880.389	ZINB	ZINB/ ZHNB
ZINB	247.099	267.228		
ZHP	2634.575	2650.678	ZHNB	
ZHNB	550.921	571.050		

Where the Vuong test returns (as in the previous cases) that both models fit equally well; the residuals and the Rootogram of both models are studied again to ensure the test result *(Figure 15)*:

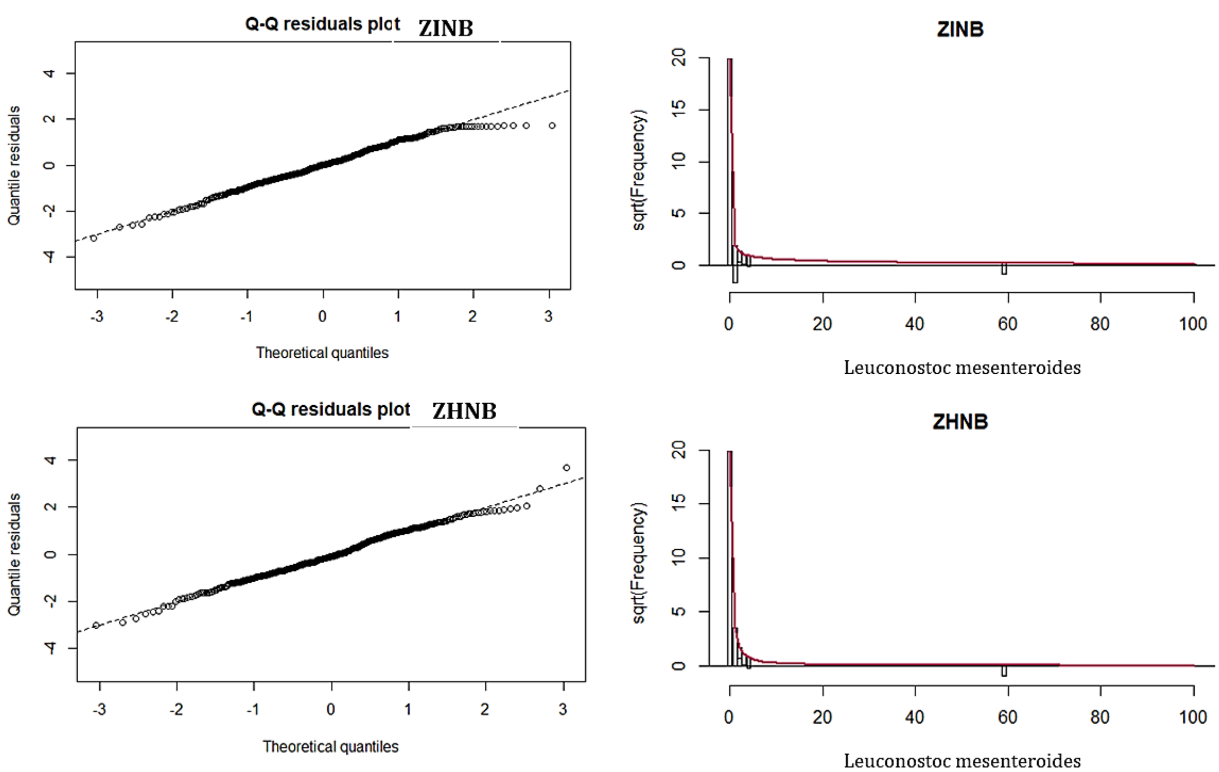


Figure 15. *Q-Q Residuals Plot and Rootogram for zero inflated Negative Binomial model and zero hurdle Negative Binomial model. In the graph on the left are the residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

Where it can be seen that for both models the residue graph is practically the same, and that they fit correctly to line in a limit between 2 and -2, indicating the normality of these. The main difference between these two graphs can be found at the far right, where the ZINB model residues are "moving away" from the line. Something that also happened in the *Figure 11*.
On the other hand, the Rootograms are different for each of the models, the curve of the Rootogram for the ZHNB model is much smoother and fits better to the observed data than in the Rootogram for ZINB.

Therefore, it is assumed that the data fits better in a Zero-Hurdle model; the result of this is *(Output 8)*:

```
Call:
hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")
Count model coefficients (truncated negbin with log link):
                      Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)           -21.569       94.701         -0.228       0.8198
exprTLR7                3.582        1.203          2.978       0.0029 **
Log(theta)            -13.012       94.700         -0.137       0.8907

Zero hurdle model coefficients (binomial with logit link):
                      Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)           -2.99132      0.23094        -12.953      <2e-16 ***
exprTLR7               0.03671      0.06136          0.598      0.55


---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 0.1451
Number of iterations in BFGS optimization: 23
Log-likelihood: -270.5 on 5 Df
```

Output 8

Interpretation: *(Zero-hurdle model)* For those subjects with minimal values of *TLR7* gene expression, the probability of having a positive count of *Leuconostoc mesenteroides* is 0.05:

$$P\left(L.\widehat{Mesenteroides} \ \geq 1|minExprTLR7\right) = \frac{e^{-2.99132}}{\left(1+e^{-2.99132}\right)} = 0.0478 \sim 0.05$$

(Positive count model) For each 1-unit increase in the *"Z-Score"*, of the *TLR7* gene expression, there will be a ($e^{3.582} = 35.945$) 36 increase in the average abundance of *Leuconostoc mesenteroides.*

The theta count confirms that the data does present over-dispersion. So the link function (taking into account *Equations 10 and 11*) would look like this:

- For the logistic regression:

$$\left(\frac{p_j}{1-p_j}\right) = e^{-2.99132}$$

- For the truncated model, there will be an increase in μ_j:

$$(\mu_j) = e^{3.582} * Expr.TLR7 = 35.945 * Expr.TLR7$$

It can be concluded that the Vuong test still gives the same results, and that the zero hurdle models give better results than the zero inflated models. For the categorical variables with more than two levels, it is still not possible to have results due to the high percentage of zeros that the bacteria still present.

IV. Bacteria with 94.9% ~ 70% zeros

Only one bacterium will be studied that has enough positive counts to carry out the models with categorical variables of more than two levels, without returning lost values in the results of the models: *Streptococcus mitis.*

$$\mathbb{E}_{S.Mitis} = 0.963, \qquad \mathbb{V}ar_{S.Mitis} = 85.948, \qquad \mathbb{CV}_{S.Mitis} = 962.65\%$$

As in the other cases that have been shown, the variance is much greater than the expect- value and the coefficient of variation is still extremely high, indicating a clear dispersion of the data. The first covariate with which the model will be made is *"Tumor_Stage"*, which has 4 levels and none of these contain only zeros, so the model can be carried correctly *(Table 11)*:

Table 11. *S. Mitis bacterium with Stage Tumour. This table shows each of the categories of this covariate and the number of zeros and counts that each of them contains.*

S. Mitis	***Stage i***	***Stage ii***	***Stage iii***	***Stage iv***	***NA***
Zero	2	104	111	115	0
Positive count	2	30	38	29	3

A comparison between the 4 plausible models is shown below *(Table 12):*

Table 12. *Model comparison based on AIC, BIC, likelihood ratio test and Vuong test. In this table, the results for the AIC and BIC criteria and the result of the likelihood ratio test are shown for each of the above models, with Streptococcus mitis bacteria and the covariate "Tumor_stage".*

Models	*AIC*	*BIC*	*LHRT*	*Vuong test*
ZIP	2450.670	2483.199	ZINB	ZINB/ ZHNB
ZINB	827.702	864.297		
ZHP	2438.233	2470.761	ZHNB	
ZHNB	796.060	832.655		

The Vuong test returns the same result as in previous cases (ZINB/ ZHNB, both models are valid) so the residuals and Rootograms of the finalist models, ZINB vs. ZHNB, are studied, in order to select the one that best fits the data *(Figure 16):*

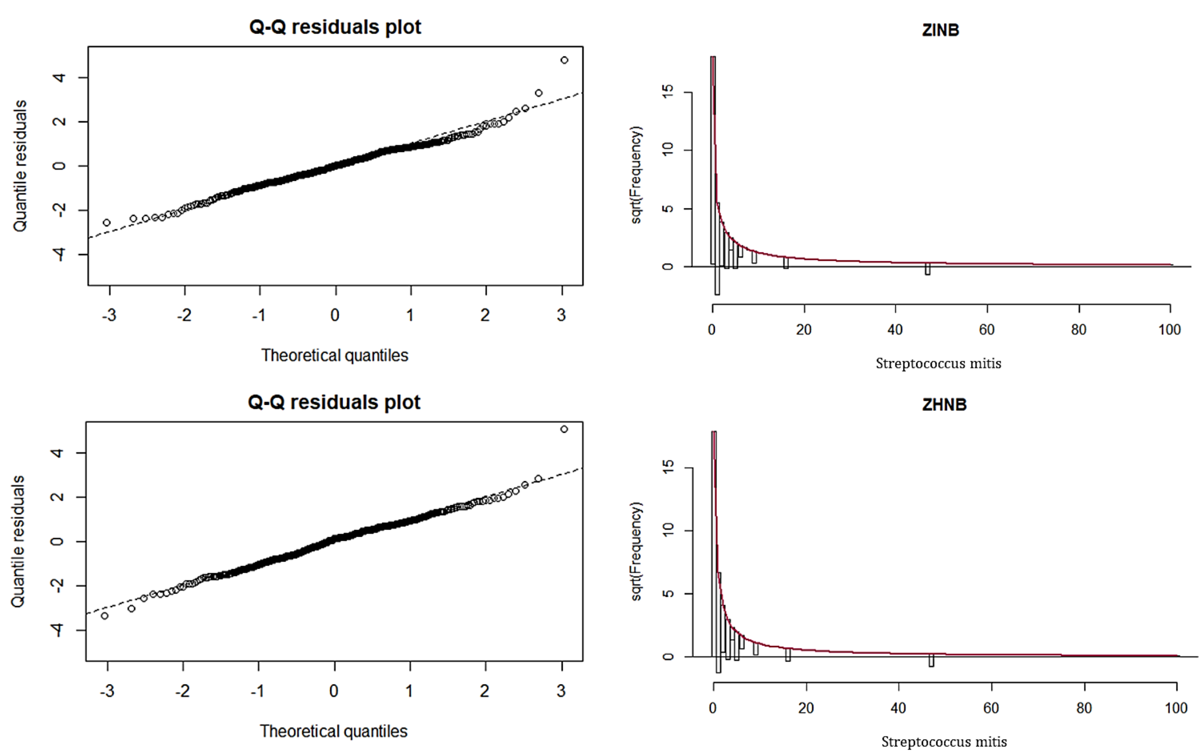


Figure 16. *Q-Q Residuals Plot and Rootogram for zero inflated Negative Binomial model and zero hurdle Negative Binomial model. In the graph on the left are the residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

In both models the residuals are very similar and fit correctly to the line between 2 and -2, thus fulfilling the hypothesis of normality of the residuals. The Rootograms of the models are different, it can be seen that the expected data (the red line) fits better to the data observed in the zero-hurdle model, so this model is selected as the final model; the results are *(Output 9 – Annexes page 16):*

There are no results that are statistically significant, the theta count confirms the over-dispersion of data.

In all the blocks/scenarios where the number of zeros and the number of positive counts were studied (regardless of the covariate used to carry out the models), the model that best fitted the zeros was the Zero Hurdle Model. This is a very important result since it will help us a lot when automating the whole database, in the articles that have been mentioned along the different sections, only one case was studied (a concrete percentage of zeros) but never with databases as big as the ones that are being applied here; when verifying that independently the percentage of zeros, the zero hurdle model works correctly it will be possible to automate all this much better.
One advantage that needs to be mentioned is that the Zero Inflated Model and Zero Hurdle models is that, when considered as two-part models, different covariates can be used in each of those parts, for example, if the formula being declared contains "|" the parts of the model are being separated and it would have to be declared which covariate is going to be applied to each part of the model, *Frame 4*:

```
f1 <- formula (Leuconostoc mesenteroides ~ exprTLR7 + offset(Offset)
          |exprTLR7)
```

Frame 4

In *"f1"* we would be stating that the covariate *"ExprTLR7"* is going to be applied in both parts of the model; this is a great advantage in multivariate models, since different covariates could be applied in different parts of the models.

6.2. Cluster without excess zeros:

In this other section, it is found the bacteria that do not present excess of zero, that it was found when making the *k-means* cluster with the ***Kraken Count database***; as it is indicated in the following scheme *(Figure 17)*, half of these do not even present zeros, and of those that present it will be analysed if they respond better to a normal GLM or if it is necessary to use more complex methods, as the previous models that have been used. In bacteria whose minimum count is not zero, these models clearly cannot be tested, since they do not present zeros.

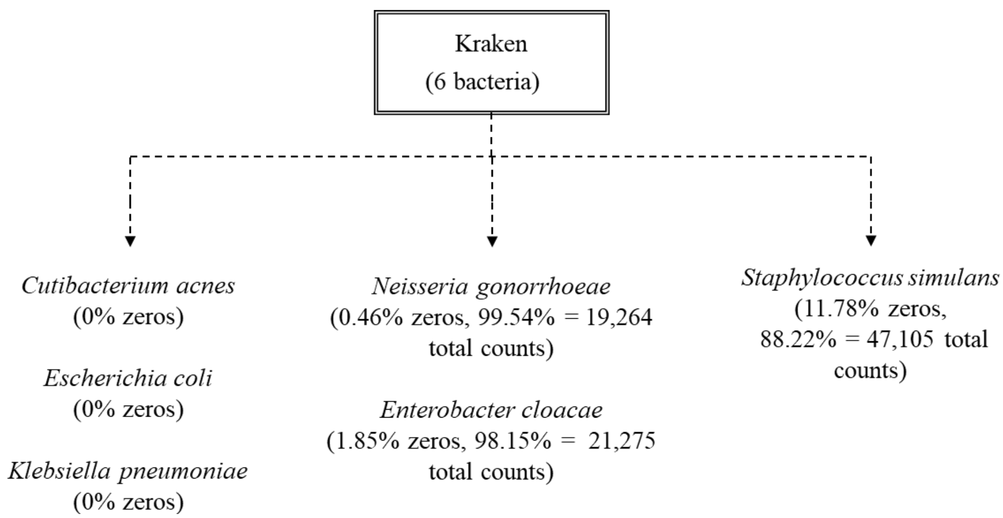


Figure 17. *Kraken diagram. This diagram shows the percentage of zeros that this database has, and the number of bacteria that are characterized by it. Graph created in Biorender.com.*

In the case of bacteria such as *Neisseria gonorrhoeae*, *Enterobacter cloacae* or *Staphylococcus simulans*, there is a very low percentage of zeros and very high positive counts. When analysing the theory behind the Zero Inflated models and the Zero Hurdle models, it has been seen that the main difference between these models is that, while the Zero Inflated Model is based on a distribution with a mass function concentrated on zero, the Zero Hurdle Model differentiates between zeros and non-zeros and fits a model for each situation.

Clearly in this situation a Zero Inflated model cannot be carried out (since there is no excess of zeros), but a Zero Hurdle model can. Therefore, to check if the Zero Hurdle model works better than a GLM we will take a bacterium that has a low percentage of zeros:

⇒ ***Case 1:*** bacterium with 8 zeros and a total of 21,275 positive counts.

Enterobacter cloacae, with 8 zeros (1.85%) and a total of 21,275 positive values,

$$\mathbb{E}_{E.\,Cloacae} = 49{,}134 \qquad \mathbb{V}ar_{E.Cloacae} = 4925,\ 153 \qquad \mathbb{CV}_{E.Cloacae} = 142.83\%$$

The variation has been found to be larger than the expected value of the data, indicating a high over-dispersion. As the CV is less than 20% it indicates a high accuracy of the data and that the average is representative of the data set, therefore it is "homogeneous". The models are built exactly as before, the *offset* and the formula are the same, *Frame 5*:

```
Offset <- log(totalReadsKraken$nonHumanReads)
fm < - formula (E.Cloacae ~ Gender + offset(Offset))
```

Frame 5

In this case it will be used as a covariate, gender, therefore the models to be studied are as follows, *Frame 6*:

```
GLMNB <- glm.nb(fm, data = metadataBCLA_RNA))
GLMP <- glm (fm, family = "poisson", data = metadataBCLA_RNA))
ZHP <- hurdle (formula = fm, dist= "poisson", data = metadataBCLA_RNA)
ZHNB <- hurdle (formula = fm, dist= "negbin", data = metadataBCLA_RNA)
```

Frame 6

The same model selection criteria that have been used in the previous blocks will be carried out *(Table 13):*

Table 13. *Model comparison based on AIC, BIC, likelihood ratio test and Vuong test. In this table, the results for the AIC and BIC criteria and the result of the likelihood ratio test are shown for each of the above models, with Enterobacter cloacae bacteria and the covariate "Gender".*

Models	*AIC*	*BIC*	*LHRT*	*Vuong test*
GLMP	49158.858	49167	GLMNB	
GLMNB	4468.375	4480.587		
ZHP	49118.059	49134.341	ZHNB	ZHNB
ZHNB	4412.751	4433.105		

Those models have lower values in the AIC and Bic criteria are the ones that uses a Negative Binomial distribution, and they are also the ones that select the Likelihood Ratio Test. The main difference, is that this is the first time that the Vuong test has selected a final model *(Output 10):*

Model 1 Class: glm.negbin

Model 2 Class: hurdle negbin

1. Variance test:
H0: Model 1 and Model 2 are indistinguishable
H1: Model 1 and Model 2 are distinguishable
w2 = 0.086, p = 2.18e-08

2. Non-nested likelihood ratio test:
H0: Model fits are equal for the focal population
H1A: Model 1 fits better than Model 2
z = -4.889, p = 1

H1B: Model 2 fits better than Model 1
z = -4.889, p = 5.077e-07

Output 10

In the *Variance test,* the *p_value* is statistically significant, so the alternative hypothesis that models are distinguishable is accepted. Since the models are distinguishable we will study which of the two models fits our data best in the *Non-nested likelihood ratio test*. We see that the hypothesis with a significant p_value is the H1B, so the Zero Hurdle Negative Binomial model (model 2), fits better than the GLM Negative Binomial (model 1). In addition, if the residuals and Rootograms of both models are compared *(Figure 18):*

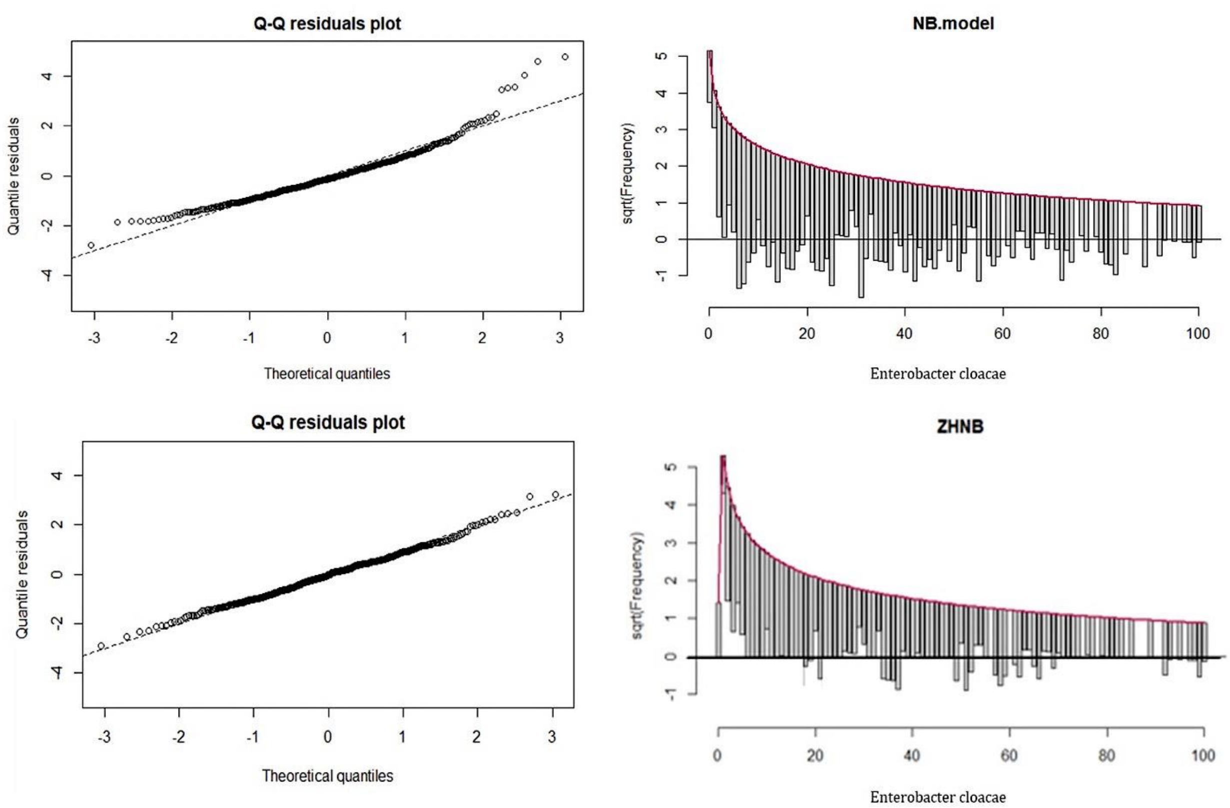


Figure 18. *Q-Q Residuals Plot and Rootogram for GLM Negative Binomial model and Zero Hurdle Negative Binomial model. In the graph on the left are the residues and, on the right, the Rootogram. The Q-Q Residuals Plot has been created with the "stats" package and the Rootogram with the "countreg" package of the RStudio software.*

In both models the residuals are very similar and fit correctly to the line between 2 and -2, thus fulfilling the hypothesis of normality of the residuals, although in the zero hurdle model negative binomial model these fit much better to the line. If we analyse the Rootograms in the first counts of the GLM. Negative Binomial model we see a mismatch (over the line), which does not occur in the Zero Hurdle Negative Binomial model, in addition, in this model all bars are on the 0 line, while in the GLM model, many are being over/under fitting. The result of the ZHNB model is *(Output 11):*

```
Call:
hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")

Count model coefficients (truncated negbin with log link):
              Estimate     Std. Error    z value     Pr(>|z|)
(Intercept)   -9.7614      0.1413        -69.062     < 2e-16 ***
Male           0.4702      0.1622          2.899     0.00374 **
Log(theta)    -0.8664      0.1041         -8.322     < 2e-16 ***

Zero hurdle model coefficients (binomial with logit link):
              Estimate     Std. Error    z value     Pr(>|z|)
(Intercept)    4.0604      0.7132        5.693       1.24e-08 ***
Male          18.5056      4477.1768     0.004       0.997
---
Theta: count = 0.4204
```

Output 11

Interpretation: *(Zero-hurdle model)* The probability (for a woman) of having a positive count is 0.985, and the probability (for a woman) of having a zero count is 0.015:

$$P\left(\widehat{E.Cloacae \geq} 1|Woman\right) = \frac{e^{4.0604}}{(1+e^{4.0604})} = 0.9845 \sim 0.985$$

The probability (for a male) of having a positive count is 1, and the probability (for a men) of having a zero count is 0, (the category is **not significant**, it is just an example of how it would be calculated):

$$P\left(\widehat{E.Cloacae \geq} 1|Male\right) = \frac{e^{4.0604+18.5056}}{(1+e^{4.0604+18.5056})} = 1$$

(Positive count model) The mean frequency of *Enterobacter cloace* for a woman with mussel invasive bladder cancer is ($e^{-9.7614} = 0.00006$), 0. The mean frequency of *Enterobacter cloace* for a male with mussel invasive bladder cancer is 1.60 times, the mean frequency of *Enterobacter cloacae* for a woman with the same type of pathology, holding the other variables constant. The theta count confirms that the data does present over-dispersion.

Therefore, the link function (taking into account *Equations 10 and 11*) would look like this:

- For the logistic regression:

$$\left(\frac{p_j}{1-p_j}\right) = e^{4.0604}$$

- For the truncated model, there is an increase in μ_j:

$$(\mu_j) = e^{-9.7614} + e^{0.4702} * Gender_{Male} = 0.01689 \ + 1.60 * Gender_{Male}$$

With this, it is demonstrated that the Zero Hurdle model works better than a GLM, with data presenting any percentage of zeros, then it will show a bacterium that does not present any 0 where a GLM has to be applied.

⇒ ***Case 2:*** bacterium without zeros.

Escherichia coli, with 0 zeros:

$$\mathbb{E}_{E.\,Coli} = 40{,}177 \qquad \mathbb{V}ar_{E.Coli} = 965{,}498 \qquad \mathbb{CV}_{E.Coli} = 77.34\%$$

As the variance is much larger than the expected value, it will directly adjust to a negative binomial distribution, as the data are clearly over-dispersed. The models are built exactly as before, the *offset* and the formula are the same, *Frame 7*:

```
Offset <- log(totalReadsKraken$nonHumanReads)
fm1 <- formula (E.Coli ~ Immune.Subtype + offset(Offset))
GLMNB <- glm. nb(fm1, data = metadataBCLA_RNA))
```

Frame 7

Using *"Immune. Subtype"* as a covariate, the result of the model is as follows *(Output 12):*

```
Call:
glm.nb(formula = f, data = metadataBCLA_RNA, init.theta = 0.9311887962, link = log)
Coefficients:
                    Estimate    Std. Error    z value    Pr(>|z|)
(Intercept)         -9.62737    0.07735       -124.471   < 2e-16 ***
Immune.SubtypeC2     0.14679    0.11127          1.319     0.18706
Immune.SubtypeC3     0.60333    0.23608          2.556     0.01060 *
Immune.SubtypeC4     0.62953    0.18658          3.374     0.00074 ***
Immune.SubtypeC6     0.19355    0.61241          0.316     0.75196
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
(Dispersion parameter for Negative Binomial(0.9312) family taken to be 1)

  Null deviance: 501.29  on 418  degrees of freedom
  Residual deviance: 483.96  on 414  degrees of freedom
    (14 observations deleted due to missingness)
  AIC: 4244
```

Output 12

Interpretation: The presence of the *C3* immune subtype increases the average *Escherichia coli* by $e^{0,60333} = 1{,}82\%$, this corresponds to a probability of:

$$p = \frac{e^{0.60333}}{1 + e^{0.60333}} = 0.646 \sim 0.65$$

And the presence of the *C4* immune subtype increases the average *Escherichia Coli* by $e^{0.62953} = 1.88\%$, this corresponds to a probability of:

$$p = \frac{e^{0.62953}}{1 + e^{0.62953}} = 0.65$$

The link function will remain (for *C3* Immune subtype and *C4* Immune subtype):

$$logit(Escherichia\ coli) = -9.62737 + 0.60333 * C3;$$

$$logit(Escherichia\ coli) = -9.62737 + 0.62953 * C4;$$

Where the value of the probability of *Escherichia coli* can be obtained with the inverse of the natural logarithm:

$$p(Eschericiha\ coli) = \frac{e^{-9.62737} + e^{0.60333} * C3}{1 + e^{-9.62737} + e^{0.60333} * C3};$$

$$p(Eschericiha\ coli) = \frac{e^{-9.62737} + e^{0.62953} * C4}{1 + e^{-9.62737} + e^{0.66953} * C4};$$

6.2.1. Example with *DESeq2*:

As explained in Section 4.3. *Models for counting data without excess of zeros,* the example that will be shown below is within the *DESeq2* package, which can be installed through *"BiocManager":* with the variance stabilization technique this package has the capacity to model microbial data with overdispersion and without excess of zeros (this method supports bacteria containing up to 15%~20% zeros), from the sequencing of the 16S RNA-seq. *DESeq2* requires content data in the form of an integer value matrix as input data. These tables of contents are generated from RNA-Seq or other high throughput sequencing experiments.

The example datasets consist of two parts: OTU-table, where the cluster will be used without excess zeros, and a meta-table, where metadata variables will be used. Both numerical and categorical covariates can be used, the main disadvantage of this method is that covariates containing missing values cannot be used.

To explain how the analysis would be carried out, everything will be explained step by step, adding in each of these the code in RStudio:

1. **Create the count table:** the *DESeq2* needs count data in the form of a rectangular table (matrix) of integer values as input data. The table cell in the *i* row and the *j* column of the table tells how many reads have been mapped to taxon *i* in sample *j*. Once it has the matrix, the *DESeqDataSetFromMatrix()* function is used to create a *DESeq2* object, *Frame 8*.

```
matrix_clst1<-as(clst1, "matrix")
matrix_clst1<-(t(matrix_clst1))
```

Frame 8

2. **Select the metadata variable for the analysis:** the metadata consists mainly of the sample information of our interest, here one selects a categorical or numerical variable that does not contain NA, for example *"sample.type"* from the metadataBCLA_RNA database, *Frame 9*:

```
group_tumor<- metadata$Sample.Type
head(group_tumor)
factor_tumor <- data.frame(row.names=colnames(matrix_clst1),gro
up=
group_tumor)
head(factor_tumor,3)
```

Frame 9

3. **Build the DESeq2 Object:** the object class used by *DESeq2* to store the read counts and estimated values during statistical analysis is the *DESeqDataSet*, which will generally be represented in the code as a "*dds"* object. A *DESeqDataSet* object must have an associated design formula; the design formula expresses the variables that will be used in the modeling. The formula must be a tilde (~) followed by the variables to be used.
The *DESeqDataSetFromMatrix* function can be used if you already have a read count matrix prepared, with the count matrix *matrix_clst1*, and the *colData* sample information, we can build a *DESeqDataSet*, *Frame 10*:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
install.packages("BiocManager");
BiocManager::install("DESeq2");library(DESeq2)
dds <- DESeqDataSetFromMatrix (countData = matrix_clst1,
colData = factor_tumor, design = ~ group)
```

Frame 10

In *countData* you define the matrix that contains the count data of the taxon, in *colData* the object that has been created in step 2 that refers to the subject and the category of "*sample_type"* that it has. With design (which refers to the formula) you are asked to make a comparison by groups (*Solid Tissue Normal vs. Primary Tumor*). If for example a numerical variable had been selected without missing values (e.g. Leukocyte. Fraction), the code would be as follows, *Frame 11*:

```
matrix_clst1<-as(clst1, "matrix")
matrix_clst1<-(t(matrix_clst1))
V.M <- metadata$Leukocyte.Fraction
data <- data.frame(row.names=colnames(matrix_clst1),var = V.M)
dds <- DESeqDataSetFromMatrix(countData = matrix_clst1, colData
 = data, design = ~ var)
```

Frame 11

4. **Estimate size factors:** *DESeq2* uses the "median ratio method" described in Anders and Huber (2010) to estimate the size factors. It first defines a virtual reference sample by taking the median of each taxa values across samples and then computes size factors as the median of ratios of each sample to the reference sample (Anders and Huber 2010). An *offset* is built in the statistical model of *DESeq2*, *Frame 12*. The estimated size factors can be accessed using the accessor function *sizeFactors()*.

```
dds <- estimateSizeFactors(dds)
sizeFactors(dds)
```

Frame 12

5. **Estimate the Overdispersion:** the first task in the analysis of abundance microbiome data is to estimate the dispersion parameter for each taxon. When a negative binomial model is fitted, the variability between replicates is modelled by the dispersion parameter. The function *estimateDispersions()* are used to estimate the dispersion parameters, *Frame 13*:

```
print(dds<- estimateDispersions(dds))
```

Frame 13

6. **Extract the Results Table:** before carrying out the model, we make sure that the factor variable is defined correctly: *"Solid Tissue Normal"* is the first level in the condition factor, so that the default log2 fold changes are calculated as *"Primary Tumor"* over *"Solid Tissue Normal"*, *Frame 14*:

```
dds$group <- factor(dds$group, levels = c("Solid Tissue Normal",
       "Primary Tumor"))
```

Frame 14

DESeq2 conducts the differential expression analysis based on the Negative Binomial distribution. The workflow it carries out is as follows: *(i)*. estimation of size factors, *(ii)*. estimation of dispersion, (3). Negative Binomial GLM fitting and Wald statistics. After the

DESeq2 function returns a *DESeqDataSet* object, results tables (log2 fold changes and p-values) can be generated using the *results()* function, *Frame 15*:

```
dds <- DESeq(dds)
(res <- results(dds))
```

Frame 15

The result of this method is as follows *(Output 13 – Annexes page 16):*

The main results are in *baseMean*, *log2FoldChange* and *p-value*: *baseMean*, is the average of the normalized count values, dividing by size factors, taken over all samples. The remaining four columns refer to a specific contrast: the comparison of the levels *Primary Tumor vs. Solid Tissue Normal* of the factor variable group. The column *log2FoldChange* is the effect size estimate. It tells us how much the OTU's abundance seems to be different due to group with *Primary Tumor* in comparison to *Solid Tissue Normal*. This value is reported on a logarithmic scale to base 2. Those bacteria that are statistically significant, that is, that have a pvalue < 0.05, are: *Cutibacterium acnes, Neisseria gonorrhoeae, Enterobacter cloacae and Klebsiella pneumoniae.*

If the GLM models are carried out for those bacteria without zeros and the Zero Hurdle Model for the rest, the same statistically significant bacteria are obtained as with the *DESeq2* method. The main difference between applying the regression models and the method proposed by *DESeq2*, is that the different model selection methods proposed (AIC, BIC, Likelihood Ratio Test and Vuong test) cannot be applied to analyse which one best fits the original data. The main advantage is the output that this method returns, it is not necessary to select one by one the bacteria to carry out all the models, but it automatically analyses all the bacteria in the database and carries out all the models.

7. Discussion and Final Conclusion:

The analysis of microbiome data with excess zeros and overdispersion can be carried out with different methods: adapting regression models for count data with these characteristics, applying a normalization, and adjusting them to a Gaussian distribution, or using a compositional methodology.

Current RNA-Seq-based standardization methods that have been adapted for microbiome data do not take into account the unique characteristics of microbiome data. L. Chen et al. in 2018 proposed a standardization method for this type of count data (with excess zeros and overdispersion) based on geometric mean of pairwise ratios. Normalization is especially critical when library size is a confounder that correlates with the variable of interest. An inappropriate method of normalization can reduce statistical power by introducing unwanted variations or result in falsely discovered characteristics.

The microbiome data can be treated as compositional because the information contained in the abundance tables is relative. In a microbiome abundance table, the total number of counts per sample is highly variable and is limited by the maximum number of DNA readings that can be provided by the sequencer (Calle, 2019). However, our low biomass tumor associated bacteria do not clearly follow this phenomenon, as the presence of bacteria is constrained, not by the sequencing depth, but by other factors as performance of purification techniques or the capture of the messenger RNAs, or maybe it is due to luck, or to contamination in a further step. So, we should review if this assumption is accomplished in tumor low biomass samples.

It was decided not to carry out either of these two methods because compositional methodology requires zeros to be imputed rather than modelled, and with standardisation the zeros would be transformed. When applying regression models, such as the Zero Hurdle Model or Zero Inflated Models, these directly model and adjust them to a specific distribution, it is not necessary to either impute or transform them, and this was the main objective of the project: "the management of zero values".

With these models I have demonstrated that the data present a clear over-dispersion, since they adjust much better to a Negative Binomial distribution than to a Poisson distribution and that they are better modelled with the Zero Hurdle Model, than when modelling on the one hand the positive counts and on the other hand the zeros (and that they are models that are not based on a distribution with a mass function concentrated on zero, like the Zero Inflated Model) work with low percentages of zeros. The fact that they are better modelled with the Zero Hurdle Negative Binomial model means that the three types of zeros defined by Kaul et al., in 2017, are in the ***Kraken Counts database***.

By analysing all possible situations with different percentages of zeros, it was shown that the Zero Hurdle Model works better than the other options regardless of the number of zeros the bacteria have, so the database was automated. Those bacteria without zeros were modelled with the GLM Negative Binomial and those with zeros, with the Zero Hurdle Negative Binomials Models (all these results are in the appendices), using as covariates the variables of *Table 3*. Previously, having eliminated the cases with a percentage < 99% of zeros.

All articles, such as those by Chen, P et al (2018) study the behaviour of a single situation with a sufficient number of zeros to be able to declare it as excess zeros. None of these articles apply these models (neither the Zero Inflated Model nor the Zero Hurdle Model) to a large database, nor do they check what happens with those cases with "low" percentages of zeros. Therefore, demonstrating that ZHNB work perfectly was a great advantage when automating the database analysis. So, the workflow developed in *Figure 6*, would be updated to the next one *(Figure 19)*:

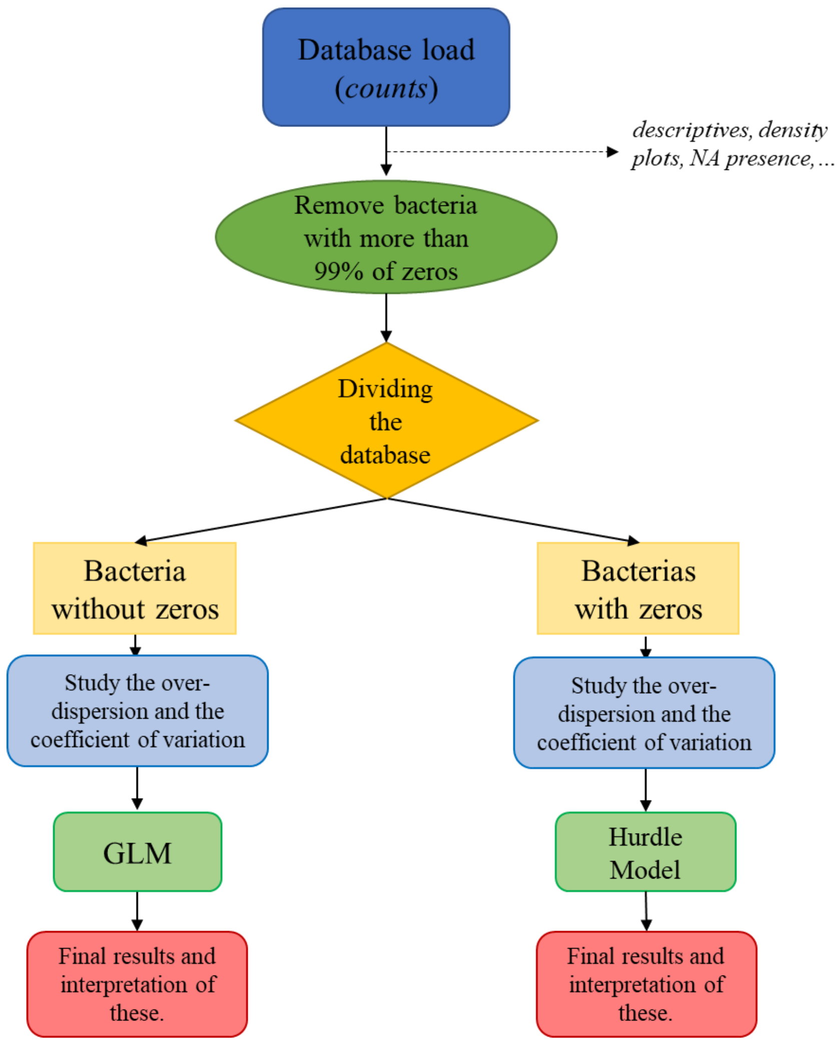


Figure 19. *Updated workflow for Figure 5. Note that any percentage of zeros is better modelled with a Zero Hurdle Model than with other models, the database is divided into two groups: bacteria that have zeros and bacteria without zeros. Those bacteria that have no zeros are modelled with a GLM, and those that have zeros are modelled with a Zero Hurdle Model.*

All the results obtained (applying the workflow developed in *Graph 18*) can be found by accessing the following QR code, which range from page one to page 425:

One thing that needs to be pointed out is that these models can not only be applied to microbiome data, but they can be applied to any kind of *omic* count data with zeros, offering a very high projection and applicability.

In addition, in recent months, regression models for zero count data have become very important and have become more widely known, due to the global pandemic experienced during 2020. Many researchers have used these types of models to model CoVid-19 data in hospitals (Ilyas et al., 2020) , nursing homes (Goldfeld, 2020), mortality (O Adepoju et al., 2020), etc.

With this I would like to emphasize the great applicability that these models have and that it is not necessary to have a high percentage of zeros to be able to carry them out. The only condition that must be met is that they must be count data and that they must have zeros, so they can be carried out in any field of research.

8. Future planes:

This Project will continue to be developed during the coming months thanks to a grant from Center for Biomedical Research Network Cancer (CIBERONC) supported by Dr. Malats, Group Leader of the laboratory of Genetic and Molecular Epidemiology Group, of the National Cancer Research Center (CNIO). The project will also be co-directed by María Dolores Alonso, who in the same way that she has been supporting and accompanying me throughout the final master's thesis, will continue to do so during this grant. She will provide me with the biological and bioinformatics vision that I lack, and with which I have carried out multiple debates on the limitations of this type of data, making me understand that not only must a methodology be taken into account, but also the biological limitations of the data. It is a real pleasure for me to have been given this opportunity, and for such professional people to accompany me.

A scientific article will be carried out describing the methodology used in this thesis and its biological impact, specifying which bacteria proved to be statistically significant and their effect on the different covariates. In addition, the standardization proposed by L. Chen et al. will be applied and it will be analysed if the data are better modelled with regression models for count data, or applying a previous standardization and adjusting it to a Gaussian distribution.

A compositional methodology is also intended to be carried out, taking into account the biological limitations of these data.

9. Bibliography:

Aitchison, J., & Kay, J. W. (2003). Possible solution of some essential zero problems in compositional data analysis. In S. Thió-Henestrosa & J. A. Martín-Fernández (Eds.), *Proceedings of CoDaWork'03, The 1st Compositional Data Analysis Workshop*. University of Girona.

Aragón, I. M., Herrera-Imbroda, B., Queipo-Ortuño, M. I., Castillo, E., Del Moral, J. S.-G., Gómez-Millán, J., Yucel, G., & Lara, M. F. (2018). The Urinary Tract Microbiome in Health and Disease. *European Urology Focus*, *4*(1), 128–138.

Arivaradarajan, P., & Misra, G. (2019). *Omics Approaches, Technologies And Applications: Integrative Approaches For Understanding OMICS Data*. Springer.

Ariza-Andraca, R., & García-Ronquillo, M. (2016). El microbioma humano. Su papel en la salud y en algunas enfermedades. *Cirugía y Cirujanos*, *84*(Supl.1), 31–35.

Bi, H., Tian, Y., Song, C., Li, J., Liu, T., Chen, Z., Chen, C., Huang, Y., & Zhang, Y. (2019). Urinary microbiota – a potential biomarker and therapeutic target for bladder cancer. *Journal of Medical Microbiology*, *68*(10), 1471–1478. https://doi.org/10.1099/jmm.0.001058

Bozdogan, H. (1987). Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. *Psychometrika*, *52*(3), 345–370. https://doi.org/10.1007/BF02294361

Calle, M. L. (2019). Statistical Analysis of Metagenomics Data. *Genomics & Informatics*, *17*(1). https://doi.org/10.5808/GI.2019.17.1.e6

Cambiaghi, A., Ferrario, M., & Masseroli, M. (2017). Analysis of metabolomic data: Tools, current strategies and future challenges for omics data integration. *Briefings in Bioinformatics*, *18*(3),

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., … Knight, R. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nature Methods*,

Chen, E. Z., & Li, H. (2016). A two-part mixed-effects model for analyzing longitudinal microbiome compositional data. *Bioinformatics (Oxford, England)*, *32*(17), 2611–2617.

Chen, J., Bushman, F. D., Lewis, J. D., Wu, G. D., & Li, H. (2013). Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, *14*(2), 244–258. https://doi.org/10.1093/biostatistics/kxs038

Fisher, R. A. (1941). The Negative Binomial Distribution. *Annals of Eugenics*, *11*(1), 182–187.

Gagnaire, A., Nadel, B., Raoult, D., Neefjes, J., & Gorvel, J.-P. (2017). Collateral damage: Insights into bacterial mechanisms that predispose host cells to cancer. *Nature Reviews. Microbiology*, *15*(2), Gilbert, J. A., Blaser, M. J., Caporaso, J. G., Jansson, J. K., Lynch, S. V., & Knight, R. (2018). Current understanding of the human microbiome. *Nature Medicine*, *24*(4), 392–400.

Goldfeld, K. (n.d.). *A hurdle model for COVID-19 infections in nursing homes*. ouR data generation. Retrieved 27 September 2020, from https://www.rdatagen.net/post/a-hurdle-model-for-covid-19-infections-in-nursing-homes-sample-size-considerations/

Group, T. N. H. W., Peterson, J., Garges, S., Giovanni, M., McInnes, P., Wang, L., Schloss, J. A., Bonazzi, V., McEwen, J. E., Wetterstrand, K. A., Deal, C., Baker, C. C., Francesco, V. D., Howcroft, T. K., Karp, R. W., Lunsford, R. D., Wellington, C. R., Belachew, T., Wright, M., … Guyer, M. (2009). The NIH Human Microbiome Project. *Genome Research*, *19*(12), 2317–2323.

HE, H., TANG, W., WANG, W., & CRITS-CHRISTOPH, P. (2014). Structural zeroes and zero-inflated models. *Shanghai Archives of Psychiatry*, *26*(4), 236–242. https://doi.org/10.3969/j.issn.1002-0829.2014.04.008

Hilbe, J. M. (2017). The statistical analysis of count data / El análisis estadístico de los datos de recuento. *Culture and Education*, *29*(3), 409–460. https://doi.org/10.1080/11356405.2017.1368162

Hu, M.-C., Pavlicova, M., & Nunes, E. V. (2011). Zero-inflated and Hurdle Models of Count Data with Extra Zeros: Examples from an HIV-Risk Reduction Intervention Trial. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 367–375. https://doi.org/10.3109/00952990.2011.597280

Hugenholtz, P., & Tyson, G. W. (2008). Metagenomics. *Nature*, *455*(7212), 481–483. https://doi.org/10.1038/455481a

Hutchison, C. A. (2007). DNA sequencing: Bench to bedside and beyond. *Nucleic Acids Research*, *35*(18), 6227–6237. https://doi.org/10.1093/nar/gkm688

Ilyas, S., Srivastava, R. R., & Kim, H. (2020). Disinfection technology and strategies for COVID-19 hospital and bio-medical waste management. *Science of The Total Environment*, *749*, 141652. https://doi.org/10.1016/j.scitotenv.2020.141652

Inouye, D., Yang, E., Allen, G., & Ravikumar, P. (2017). A Review of Multivariate Distributions for Count Data Derived from the Poisson Distribution. *Wiley Interdisciplinary Reviews. Computational Statistics*, *9*(3). https://doi.org/10.1002/wics.1398

Kaul, A., Mandal, S., Davidov, O., & Peddada, S. D. (2017a). Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02114

Kaul, A., Mandal, S., Davidov, O., & Peddada, S. D. (2017b). Analysis of Microbiome Data in the Presence of Excess Zeros. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.02114

Kent, J. T. (1982). Robust properties of likelihood ratio tests. *Biometrika*, *69*(1), 19–27. https://doi.org/10.1093/biomet/69.1.19

La Rosa, P. S., Zhou, Y., Sodergren, E., Weinstock, G., & Shannon, W. D. (2015). Chapter 6—Hypothesis Testing of Metagenomic Data. In J. Izard & M. C. Rivera (Eds.), *Metagenomics for Microbiology* (pp. 81–96). Academic Press. https://doi.org/10.1016/B978-0-12-410472-3.00006

Martin, T. G., Wintle, B. A., Rhodes, J. R., Kuhnert, P. M., Field, S. A., Low-Choy, S. J., Tyre, A. J., & Possingham, H. P. (2005). Zero tolerance ecology: Improving ecological inference by modelling the source of zero observations. *Ecology Letters*, *8*(11), 1235–1246. https://doi.org/10.1111/j.1461-0248.2005.00826.x

Martín-Fernández, J.-A., Hron, K., Templ, M., Filzmoser, P., & Palarea-Albaladejo, J. (2014). Bayesian-multiplicative treatment of count zeros in compositional data sets: *Statistical Modelling*. https://doi.org/10.1177/1471082X14535524

McMurdie, P. J., & Holmes, S. (2014). Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. *PLOS Computational Biology*, *10*(4), e1003531. https://doi.org/10.1371/journal.pcbi.1003531

MetaHIT Consortium (Metagenomics of the Human Intestinal Tract consortium)—Wellcome Sanger Institute. (n.d.). Retrieved 18 March 2020, from https://www.sanger.ac.uk/resources/downloads/bacteria/metahit/

Min, Y., & Agresti, A. (2005). Random effect models for repeated measures of zero-inflated count data. *Statistical Modelling: An International Journal*, *5*(1), 1–19. https://doi.org/10.1191/1471082X05st084oa

Mullahy, J. (1986). Specification and testing of some modified count data models. *Journal of Econometrics*, *33*(3), 341–365. https://doi.org/10.1016/0304-4076(86)90002-3

Ohadian Moghadam, S., & Nowroozi, M. R. (2019). Toll-like receptors: The role in bladder cancer development, progression and immunotherapy. *Scandinavian Journal of Immunology*, *90*(6), e12818. https://doi.org/10.1111/sji.12818

Omenn, G. S., Lane, L., Lundberg, E. K., Beavis, R. C., Nesvizhskii, A. I., & Deutsch, E. W. (2015). Metrics for the Human Proteome Project 2015: Progress on the Human Proteome and Guidelines for High-Confidence Protein Identification. *Journal of Proteome Research*, *14*(9), 3452–3460. https://doi.org/10.1021/acs.jproteome.5b00499

Oresta, B., Hurle, R., Lazzeri, M., Frego, N., Saita, A., Faccani, C., Fasulo, V., Casale, P., Pozzi, C., Guazzoni, G. F., & Rescigno, M. (2020). Characterization of the urinary microbiota in bladder cancer patients. *Journal of Clinical Oncology*, *38*(6_suppl), 535–535. https://doi.org/10.1200/JCO.2020.38.6_suppl.535

Paulson, J. N., Stine, O. C., Bravo, H. C., & Pop, M. (2013). Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, *10*(12), 1200–1202. https://doi.org/10.1038/nmeth.2658

Requena, T., & Velasco, M. (2019). Microbioma humano en la salud y la enfermedad. *Revista Clínica Española*. https://doi.org/10.1016/j.rce.2019.07.004

Sanli, O., Dobruch, J., Knowles, M. A., Burger, M., Alemozaffar, M., Nielsen, M. E., & Lotan, Y. (2017). Bladder cancer. *Nature Reviews. Disease Primers*, *3*, 17022. https://doi.org/10.1038/nrdp.2017.22

Schenker, N., & Taylor, J. M. G. (1996). Partially parametric techniques for multiple imputation. *Computational Statistics & Data Analysis*, *22*(4), 425–446. https://doi.org/10.1016/0167-9473(95)00057-7

Schwarz, G. (1978). Estimating the Dimension of a Model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Sommer, F., Anderson, J. M., Bharti, R., Raes, J., & Rosenstiel, P. (2017). The resilience of the intestinal microbiota influences health and disease. *Nature Reviews Microbiology*, *15*(10), 630–638. https://doi.org/10.1038/nrmicro.2017.58

The Human Microbiota in Health and Disease | Elsevier Enhanced Reader. (n.d.). https://doi.org/10.1016/J.ENG.2017.01.008

Thorsson, V., Gibbs, D. L., Brown, S. D., Wolf, D., Bortone, D. S., Ou Yang, T.-H., Porta-Pardo, E., Gao, G. F., Plaisier, C. L., Eddy, J. A., Ziv, E., Culhane, A. C., Paull, E. O., Sivakumar, I. K. A., Gentles, A. J., Malhotra, R., Farshidfar, F., Colaprico, A., Parker, J. S., … Shmulevich, I. (2018). The Immune Landscape of Cancer. *Immunity*, *48*(4), 812-830.e14. https://doi.org/10.1016/j.immuni.2018.03.023

Tsilimigras, M. C. B., & Fodor, A. A. (2016). Compositional data analysis of the microbiome: Fundamentals, tools, and challenges. *Annals of Epidemiology*, *26*(5), 330–335. https://doi.org/10.1016/j.annepidem.2016.03.002

Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R., & Gordon, J. I. (2007). The human microbiome project. *Nature*, *449*(7164), 804–810. https://doi.org/10.1038/nature06244

Unger-Saldaña, K., Miranda, A., Zarco-Espinosa, G., Mainero-Ratchelous, F., Bargalló-Rocha, E., & Lázaro-León, J. M. (2015). Health system delay and its effect on clinical stage of breast cancer: Multicenter study. *Cancer*, *121*(13), 2198–2206. https://doi.org/10.1002/cncr.29331

van den Boogaart, K. G., & Tolosana-Delgado, R. (2008). "compositions": A unified R package to analyze compositional data. *Computers & Geosciences*, *34*(4), 320–338. https://doi.org/10.1016/j.cageo.2006.11.017

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., … Zhu, X. (2001). The Sequence of the Human Genome. *Science*, *291*(5507), 1304–1351. https://doi.org/10.1126/science.1058040

Wang, T., Cai, G., Qiu, Y., Fei, N., Zhang, M., Pang, X., Jia, W., Cai, S., & Zhao, L. (2012). Structural segregation of gut microbiota between colorectal cancer patients and healthy volunteers. *The ISME Journal*, *6*(2), 320–329. https://doi.org/10.1038/ismej.2011.109

Wu, P., Zhang, G., Zhao, J., Chen, J., Chen, Y., Huang, W., Zhong, J., & Zeng, J. (2018). Profiling the Urinary Microbiota in Male Patients With Bladder Cancer in China. *Frontiers in Cellular and Infection Microbiology*, *8*. https://doi.org/10.3389/fcimb.2018.00167

Xia, Y., & Sun, J. (2017a). Hypothesis Testing and Statistical Analysis of Microbiome. *Genes & Diseases*, *4*(3), 138–148. https://doi.org/10.1016/j.gendis.2017.06.001

Xia, Y., & Sun, J. (2017b). Hypothesis testing and statistical analysis of microbiome. *Genes & Diseases*, *4*(3), 138–148. https://doi.org/10.1016/j.gendis.2017.06.001

Xia, Y., Sun, J., & Chen, D.-G. (2018). Introductory Overview of Statistical Analysis of Microbiome Data. In Y. Xia, J. Sun, & D.-G. Chen (Eds.), *Statistical Analysis of Microbiome Data with R* (pp. 43–75). Springer. https://doi.org/10.1007/978-981-13-1534-3_3

Xu, L., Paterson, A. D., Turpin, W., & Xu, W. (2015). Assessment and Selection of Competing Models for Zero-Inflated Microbiome Data. *PLOS ONE*, *10*(7), e0129606. https://doi.org/10.1371/journal.pone.0129606

Xu, W., Yang, L., Lee, P., Huang, W. C., Nossa, C., Ma, Y., Deng, F.-M., Zhou, M., Melamed, J., & Pei, Z. (2014). Mini-review: Perspective of the microbiome in the pathogenesis of urothelial carcinoma. *American Journal of Clinical and Experimental Urology*, *2*(1), 57–61.

Yin, X., & Hilafu, H. (2015). Sequential sufficient dimension reduction for large p, small n problems. *Journal of the Royal Statistical Society Series B*, *77*(4), 879–892.

Zuur, A. F., Ieno, E. N., Walker, N. J., Saveliev, A. A., & Smith, G. M. (2009). GLM and GAM for Count Data. In A. F. Zuur, E. N. Ieno, N. Walker, A. A. Saveliev, & G. M. Smith (Eds.), *Mixed effects models and extensions in ecology with R* (pp. 209–243). Springer. https://doi.org/10.1007/978-0-387-87458-6_9

SPANISH NATIONAL CANCER RESEARCH CENTER

FACULTY OF STATISTICAL STUDIES

BIOSTATISTICS MASTER

Master's thesis: Annexes.

Study of the distribution and behaviour of the "0" values in large *omic* data arrays.

Author: Helena Fidalgo Gómez

Co-tutors: Dra. Núria Malats Riera

Dra. Mª Teresa Pérez Pérez

María Dolores Alonso Guirado

Madrid, 2020

SPANISH NATIONAL CANCER RESEARCH CENTER

FACULTY OF STATISTICAL STUDIES

BIOSTATISTICS MASTER

Master's thesis: Annexes.

Study of the distribution and behaviour of the "0" values in large *omic* data arrays.

Author: Helena Fidalgo Gómez

Co-tutors: Dra. Núria Malats Riera

Dra. Mª Teresa Pérez Pérez

María Dolores Alonso Guirado

Dra. Núria Malats Riera

Dra. Mª Teresa Pérez Pérez

Helena Fidalgo Gómez

María Dolores Alonso Guirado

Madrid, 2020

Index of Annexes:

1. Code applied throughout the final master's thesis:

Throughout the work different methods have been exposed and various graphics have been explained, all created from the RStudio software. In this section, all the packages and functions that have been used will be shown, and all the steps that have been followed to carry out the project will be described:

I. *Load and adapt the database:* as mentioned above, the database to be used is Kraken, so it is the first one to be loaded and adapted for future models. It is also shown how the Elbow and Silhouette method was carried out to obtain the optimal number of clusters and the k-means clustering.

```
kraken <- read.delim("C:/Users/hfidalgo/Desktop/Dataset/kraken_coun
ts_sorted.tsv", row.names=1)
data <- kraken[,1:433]
dim(data)
#Dertermining And Visualizing The Optimal Number Of Clusters:
library(ggplot2)
library(factoextra)
set.seed(2119)
elbow <- fviz_nbclust(Bacterias_en_filas, kmeans, method = "wss") +
labs(subtitle = "Elbow method")
silhoutte <- fviz_nbclust(data, kmeans, method = "silhouette",
print.summary = TRUE) + labs(subtitle = "Silhouette method")
print(silhoutte)
optimal <- silhoutte$data
print(optimal <- as.data.frame(optimal))
library(dplyr)
optimal2 <- optimal %>% filter(y == max(y)); as.numeric(optimal2[,1
])
# Optimal2 contains the optimal number of clusters that have been ca
lculated with the Elbow and Silhouette method. Optimal2 is included
in the "eclust" function to obtain the k-means clustering:
library(NbClust)
B.Kmeans <- eclust(data, FUNcluster = "kmeans", k= as.numeric(optim
al2[,1]), k.max = as.numeric(optimal2[,1]), stand = T, graph = T, h
c_metric = "euclidean", hc_method = "ward.D2", gap_maxSE = list(met
hods= "firstSEmax", SE.factor = 1), nboot = 100, verbose = TRUE, se
ed = 2119)
fviz_silhouette(B.Kmeans)
B.Kmeans$nbclust #Number of clusters calculated.
table(B.Kmeans$cluster)
nclust <- B.Kmeans$cluster
typeof(nclust)
nclust <- as.matrix(nclust)
nclust <- as.data.frame(nclust)
nclust <- rename(nclust, factor.cluster = V1)
#This new column has been created, and it contains two different ca
tegories: 1 and 2. Bacteria belonging to cluster 1 have category 1
and bacteria belonging to cluster 2 have category 2. This column is
added to the kraken database, so that later on the database can be
divided into two different groups:
head(kraken)
```

```
head(nclust)
kraken.f <- merge(kraken, nclust, by=0, all=TRUE)
#We separate the databases according to the cluster they belong to:
clst1 <- subset(kraken.f, factor.cluster=="1")
head(clst1);dim(clst1)
clst2 <- subset(kraken.f, factor.cluster=="2")
head(clst2,29);dim(clst2)
```

II. *Loading of the following database and selection of variables:* the next database to be used will be *metadataBCLA_RNA*, as explained at the beginning of section 5 of the project *(5. Application to a real database, page 22)*, only certain variables will be selected and the variable BMI will be imputed, using the function "mice" and as "predictive mean matching" methodology.

```
metadataBLCA_RNA <- read.delim("C:/Users/hfidalgo/Desktop/Datasets/
metadataBLCA_RNA2019-02-20_TLRs_Thorsson_clinical_exposure.tsv")
library(dplyr)
library(tidyverse)
library(tibble)
# Selection of the numerical variables:
metadata = metadataBLCA_RNA %>% select(File.Name,exprIL1B, exprTLR7
,exprNOD1, Leukocyte.Fraction,SNV.Neoantigens,bmi)
metadata$File.Name <- gsub("_gdc_realn_rehead.bam", "", metadata$Fi
le.Name)
has_rownames(metadata)
metadata <- column_to_rownames(metadata, var = "File.Name")
metadata = metadata[order(row.names(metadata)),]
head(metadata)
#For the BMI column: before imputing the variable BMI, the values "-
-" are replaced by missing values, then the imputation is carried ou
t using the predictive mean matching method (method = "pmm").
metadata$bmi <- as.character(metadata$bmi)
metadata$bmi <- as.numeric(metadata$bmi) #the values "--" have been
replaced by NA
library(mice)
metadata.imp <- mice(metadata, method = "pmm", seed=500)
densityplot (metadata_final.imp, ~ bmi|.imp)
metadata.imp <- complete(metadata.imp,1)
View(metadata.imp)
metadata$File.Name <- gsub("_gdc_realn_rehead.bam", "", metadata$Fi
le.Name)
has_rownames(metadata)
metadata <- column_to_rownames(metadata, var = "File.Name")
#We order both databases, as both databases have the same subject ID
, which are found in the row.names, so that ordering both databases
by subject ID is sufficient.
metadata = metadata[order(row.names(metadata)),]
metadata.imp = metadata.imp[order(row.names(metadata.imp)),]
#We replace the original BMI variable by the imputed BMI variable
(BMI.imp):
metadata$bmi <- NULL
metadata$bmi.imp <- metadata.imp$bmi
# With the function "cut" the variable is categorized in: "Under", "
Normal", "Overweight" and "Obese".
```

```
metadata$bmi.imp <- cut(metadata$bmi.imp, breaks = c(min(metadata$b
mi.imp), 18.5, 25, 30, max(metadata$bmi.imp)), labels = c("Under",
"Normal", "Overweight", "Obese"))
table(metadata$bmi.imp)
# The rest of the remaining variables are included:
metadata$Immune.Subtype <- metadata$Immune.Subtype
metadata$tumor_stage <- metadata$tumor_stage
metadata$tumor_stage = levels(metadata_final$tumor_stage)[levels(me
tadata_final$tumor_stage)=="not reported"] <- NA
```

III. *Sort the clst1 and clst2 databases and eliminate those bacteria with more than 99% of zeros:* the databases that have been created in the first section are ordered, and the number of zeros in each one of them is also studied: those with 99% of zeros are eliminated from the database, since it is not possible to obtain reliable results with such a high percentage of zeros.

```
#A. Clst1:
library(tibble)
library(dplyr)
head(select(clst1, 1:5))
clst1 <- rename(clst1, tax.ID = Row.names)
has_rownames(clst1)
clst1 <- remove_rownames(clst1)
clst1 <- column_to_rownames(clst1, var = "species")
library(dplyr)
clst1 <- clst1 %>%
           select(-1, -(434:435))
clst1 <- as.data.frame(t(clst1))
rownames(clst1) <- sub("X", "", rownames(clst1))
head(clst1)

#B. Clst2:
clst2 <- rename(clst2, tax.ID = Row.names)
has_rownames(clst2)
clst2 <- remove_rownames(clst2)
clst2 <- column_to_rownames(clst2, var = "species")
library(dplyr)
clst1 <- clst1 %>%
           select(-1, -(434:435))
clst2 <- as.data.frame(t(clst2))
head(clst2,8)
rownames(clst2) <- sub("X", "", rownames(clst2))

# It is checked that bacteria are only 0:
Especie_clst2 = unique(names(clst2))
for(i in 1:length(clst2)){
   n0 <-sum(clst2[[i]]=="0")
   print(paste("Bacterium:",Especie_clst2[i] ,"has",n0,"zeros"))
}
# It is verified that 1602 bacteria present 433 zeros (which is the
maximum that can contain), these bacteria cannot be modelled since
no model would converge. The next step will be to eliminate from the
"clst2" database those bacteria with more than 99% of zeros:
clst2 = clst2[,colSums(clst2==0, na.rm=T)/nrow(clst2)<0.99]
dim(clst2)
```

IV. *Test the different models, apply model selection criteria, obtain the residuals and the Rootogram:* this part of the code shows the packages and functions used to carry out the different models, as well as the creation of the *offset*, along with the different model selection criteria (AIC, BIC, Likelihood Ratio Test and Vuong Test), residuals and Rootogram.

```
#1). Offset:
totalReadsKraken <- read.delim("C:/Users/hfidalgo/Desktop/Bases de
datos/totalReadsKraken.tsv")
totalReadsKraken = totalReadsKraken[order(row.names(totalReadsKrake
n)),]
clst2 = clst2[order(row.names(clst2)),]
clst2$Offset <- log(totalReadsKraken$nonHumanReads)
clst1 = clst1[order(row.names(clst1)),]
clst1$Offset <- log(totalReadsKraken$nonHumanReads)

#2). Formula:
clst2 = clst2[order(row.names(clst2)),]
clst1 = clst1[order(row.names(clst1)),]
metadata = metadata[order(row.names(metadata)),]
f1 <- formula(clst2$`Leuconostoc mesenteroides` ~ metadata$exprTLR7
+ offset(clst2$Offset)|metadata $exprTLR7)
# f1 would be the same as put:
f2 <- formula(clst2$`Leuconostoc mesenteroides` ~ metadata$exprTLR7
+ offset(clst2$Offset))
# Formula for "clst1":
f3 <- formula(clst1$`Klebsiella pneumoniae` ~ metadata$exprTLR7+off
set(clst1$Offset))

#3). Models:
library(MASS)
library(pscl)
library(gee)
library(geepack)
     #-> Zero Inflated Models:
summary(ZIP <- zeroinfl(f1, dist = "poisson", link = "logit", data
= metadata))
summary(ZINB <- zeroinfl(f1, dist = "negbin", link = "logit", data
= metadata))
     #-> Zero Hurdle Models:
summary(ZHP <- hurdle(formula = f1, dist= "negbin", data=metadata))
summary(ZHNB <- hurdle(formula = f1, dist= "negbin", data=metadata)
)
     #-> GLM Models:
summary(GLMP  <- glm(f3, family = "poisson" ,data = metadata))
summary(NB.model <- glm.nb(f3,data = metadata))

#4). Model Selection Criteria:
library(lmtest)
library(nonnest2)
     #AIC and BIC criteria:
icci(ZHP,ZHNB)
     #Likelihood Ratio Test:
lrtest(ZHP, ZHNB)
     #Vuong Test:
vuongtest(ZINB, ZHNB)
```

```
#5). QQ Plots:
res <- resid(ZHNB, type = "pearson")
qqnorm(res)
qqline(res)

#6). Rootogram:
install.packages("countreg", repos="http://R-Forge.R-project.org")l
ibrary(countreg)
rootogram(ZHNB, max = 100) # fit up to count 100
```

V. *Final Results:* throughout the project it was shown that all those bacteria with at least 1 zero, fit much better to a Zero Hurdle Model than to a GLM. Therefore, two different models are applied to the Kraken database: bacteria without any zero fit a GLM negative binomial model and bacteria with at least 1 zero fit a Zero Hurdle Negative Binomial Model:

```
#The Kraken database is divided into 2 groups, one containing bacte
ria with zeros and one without zeros:
kraken <- read.delim("C:/Users/hfidalgo/Desktop/Datasets/kraken_cou
nts_sorted.tsv", row.names=1)
library(tibble)
has_rownames(kraken)
data <- remove_rownames(kraken)
data <- column_to_rownames(data, var = "species")
head(data)
data <- data[,-434]
data <- as.data.frame(t(data))
head(data,8)
rownames(data) <- sub("X", "", rownames(data))
#Bacteria are eliminated with <99% zeros:
data = data[,colSums(data==0, na.rm=T)/nrow(data)<0.99]
dim(data)

Bact.Nonzeros =  data[,colSums(data==0, na.rm=F)/nrow(data)<=0]
head(Bact.Nonzeros); dim(Bact.Nonzeros)# contains 3 bacteria

Bact.Zeros = data[,colSums(data==0, na.rm=T)/nrow(data)>0]
head(Bact.Zeros); dim(Bact.Zeros)#contains 1605 bacteria

#create the offset:
totalReadsKraken <- read.delim("C:/Users/hfidalgo/Desktop/Bases de
datos/totalReadsKraken.tsv")
totalReadsKraken = totalReadsKraken[order(row.names(totalReadsKrake
n)),]
Bact.Nonzeros = Bact.Nonzeros[order(row.names(Bact.Nonzeros)),]
Bact.Nonzeros$Offset <- log(totalReadsKraken$nonHumanReads)

Bact.Zeros = Bact.Zeros[order(row.names(Bact.Zeros)),]
Bact.Zeros$Offset <- log(totalReadsKraken$nonHumanReads)

#To carry out the models, the MetadataBCLA_RNA database is used, the
steps explained in section II are carried out.
Bact.Nonzeros = Bact.Nonzeros[order(row.names(Bact.Nonzeros)),]
Bact.Zeros = Bact.Zeros[order(row.names(Bact.Zeros)),]
metadata = metadata [order(row.names(metadata)),]
#Final results for bacteria with zeros:
```

```
getOption("max.print")
options(max.print=999999999)
# run n regressions:
dim(Bact.Zeros)
n <- 1605
my_lms <- lapply(1:n, function(x) hurdle(Bact.Zeros[,x] ~ metadata$
exprMAP3K7 + offset(Bact.Zeros$Offset),dist= "negbin" ,data = meta
data))
fvn <- data.frame(iteration=seq(1,n),t(sapply(my_lms,coefficients))
)
fvn <- column_to_rownames(fvn, var = "iteration")
fvn <- as.data.frame(fvn)
row.names(fvn) <- NULL
Bact.Zeros$Offset <- NULL
Especie = unique(names(Bact.Zeros))
for(i in 1:n){
  row.names(fvn)[i] <- Especie[i]
}
#A .txt document would be generated containing: count intercept, zer
o intercept,...
write.table(fvn, "Bacteria + ExprMAP3k7.txt")
#To download the p_values:
pv_int_count <- c()
pv_cov_count <- c()
pv_int_zero <- c()
pv_cov_zero <- c()
for(i in seq(1,n, by=1)){
     pv_int_count[[i]] = summary(my_lms[[i]])$coefficients$count[1,"
Pr(>|z|)"];
             pv <- as.data.frame(pv_int_count);
             pv = (t(pv));
     pv_cov_count[[i]] = summary(my_lms[[i]])$coefficients$count[2,"
Pr(>|z|)"];
             pv2 <- as.data.frame(pv_cov_count);
             pv2 = (t(pv2));
     pv_int_zero[[i]] = summary(my_lms[[i]])$coefficients$zero[1,"Pr
(>|z|)"];
             pv3 <- as.data.frame(pv_int_zero);
             pv3 = (t(pv3));
     pv_cov_zero[[i]] = summary(my_lms[[i]])$coefficients$zero[2,"Pr
(>|z|)"];
             pv4 <- as.data.frame(pv_cov_zero);
             pv4 = (t(pv4));
}
write.table(cbind(pv,pv2,pv3,pv4),file="Pvalues-Bacteria + ExprMAP3
k7.txt",row.names=F, col.names=c('Pv.Inter_count','Pv.Cov_count','P
v.Inter_zero','Pv.Cov_zero'))


#Final results for bacteria without zeros:
getOption("max.print")
options(max.print=999999999)
# run n regressions:
dim(Bact.Nonzeros)
n <- 3
my_lms <- lapply(1:n, function(x) glm.nb(Bact.Nonzeros[,x] ~ metada
ta $gender + offset(Bact.Nonzeros$Offset), data = metadata))
fvn <- data.frame(iteration=seq(1,n),t(sapply(my_lms,coefficients))
)
fvn <- column_to_rownames(fvn, var = "iteration")
fvn <- as.data.frame(fvn)
```

```
row.names(fvn) <- NULL
Bact.Nonzeros$Offset <- NULL
Especie = unique(names(Bact.Nonzeros))
for(i in 1:n){
  row.names(fvn)[i] <- Especie[i]
}
write.table(fvn, "Bacteria without zeros + Gender.txt")
p_values <- c()
for(i in seq(1,n, by=1)){
 p_values[[i]]=summary(my_lms[[i]])$coefficients[,"Pr(>|z|)"];
 pv <- as.data.frame(p_values);
 pv = (t(pv))
}
write.table(pv,file="Pvalues-Bacteria without zeros + Gender.txt",
row.names=F, col.names=c('Pv.Intercept','Pv.Covariable'))
```

2. Summary table of *MetadataBCLA_RNA* database:

Table showing the percentage of missing values in total, the median, the variance and the interquartile range of each variable:

Variable	***Percentage of missing values (%NA)***	***Median***	***Var***	***IQR***
exprCHUK	4.389	-0.389	1.113	1.281
exprCXCL8	4.389	-0.415	1.330	0.537
exprIFNA1	4.389	-0.405	0.434	0.253
exprIFNB1	4.389	-0.244	1.435	0.164
exprIFNG	4.389	-0.346	0.669	0.221
exprIKBKB	4.389	-0.280	1.351	1.412
exprIKBKG	4.389	-0.208	1.693	1.285
exprIL12A	4.389	-0.189	0.876	0.203
exprIL12B	4.389	-0.321	0.839	0.476
exprIL1B	4.389	-0.339	4.068	0.551
exprIL6	4.389	-0.210	1.041	0.195
exprIRAK1	4.389	-0.325	1.271	1.329
exprIRAK4	4.389	-0.061	1.718	1.401
exprMAP3K7	4.389	-0.458	1.605	1.353
exprMDK	4.389	-0.306	0.944	1.051
exprMYD88	4.389	-0.149	1.729	1.341
exprNOD1	4.389	0.008	1.884	0.862
exprNOD2	4.389	-0.371	0.850	0.557
exprSLC15A1	4.389	-0.361	1.342	0.672
exprTAB2	4.389	-0.214	1.012	1.280
exprTLR1	4.389	-0.333	0.768	0.699
exprTLR10	4.389	-0.224	0.645	0.087
exprTLR2	4.389	-0.366	0.998	0.805
exprTLR3	4.389	-0.436	0.787	0.806
exprTLR4	4.389	-0.416	0.636	0.738
exprTLR5	4.389	-0.237	1.540	1.027
exprTLR6	4.389	-0.422	0.854	0.684
exprTLR7	4.389	-0.336	5.719	0.659
exprTLR8	4.389	-0.434	0.923	0.531
exprTLR9	4.389	-0.272	0.516	0.302
exprTNF	4.389	-0.237	1.253	0.193
exprTRAF3	4.389	-0.322	1.022	1.307
exprTRAF6	4.389	-0.411	1.858	1.395
Immune Subtype	3.233	C2	-	-
TCGA Subtype	65.589	BLCA.2	-	-
Leukocyte Fraction	0	0.202	0.027	0.233
Stromal Fraction	2.540	0.4	0.050	0.371
Intratumor Heterogeneity	2.540	0.16	0.032	0.251

TIL Regional Fraction	26.097	4.967	42.778	7.139
Proliferation	3.233	0.558	0.274	0.595
Wound Healing	3.233	0.184	0.027	0.228
Macrophage Regulation	3.233	-0.3997	0.729	1.243
Lymphocyte Infiltration	3.233	-18641	1.508	1.751
IFN-gamma Response	3.233	0.1190	0.977	1.462
TGF-beta Response	3.233	0.0841	0.196	0.666
SNV Neoantigens	0.693	72	15113.706	96.751
Indel Neoantigens	8.083	29.5	5941.0518	73
Silent Mutation Rate	1.616	1.587	6.449	1.901
Nonsilent Mutation Rate	1.616	4.613	55.896	5.834
Number of Segments	0.924	168	15462.210	151
Fraction Altered	0.924	0.554	0.078	0.449
Aneuploidy Score	2.309	13	46.547	11
Homologous Recombination Defects	2.309	27	281.128	24
BCR Evenness	46.189	0.884	0.017	0.124
BCR Shannon	36.027	1.908	1.654	2.231
BCR Richness	36.027	9	1555.325	31
TCR Shannon	16.628	1.946	1.281	1.657
TCR Richness	3.233	5	370.366	13
TCR Evenness	24.942	0.981	0.002	0.042
CTA Score	3.695	2.869	2.537	2.771
Th1 Cells	3.233	-2376.544	735484.035	1143.974
Th2 Cells	3.233	447.678	307870.849	707.381
Th17 Cells	3.233	-3013.005	2527519.993	2273.981
B Cells Memory	3.233	0.020	0.005	0.074
B Cells Naive	3.233	0.002	0.004	0.040
Dendritic Cells Activated	3.233	0.026	0.004	0.068
Dendritic Cells Resting	3.233	0.003	0.001	0.021
Eosinophils	3.233	0	5.016e-05	0
Macrophages M0	3.233	0.0178	0.009	0.093
Macrophages M1	3.233	0.032	0.003	0.076
Macrophages M2	3.233	0.203	0.014	0.155
Mast Cells Activated	3.233	0	0.002	0.041
Mast Cells Resting	3.233	0.014	0.003	0.057
Monocytes	3.233	0.018	0.001	0.028
Neutrophils	3.233	0	0.001	0.006

NK Cells Activated	3.233	0.025	0.001	0.047
NK Cells Resting	3.233	0.001	0.001	0.030
Plasma Cells	3.233	0.030	0.004	0.066
T Cells CD4 Memory Activated	3.233	0	0.001	0.005
T Cells CD4 Memory Resting	3.233	0.012	0.003	0.077
T Cells CD4 Naive	3.233	0	0.005	0.020
T Cells CD8	3.233	0.109	0.007	0.104
T Cells Follicular Helper	3.233	0.0839	0.004	0.074
T Cells gamma delta	3.233	0	2.209e-05	0
T Cells Regulatory Tregs	3.233	0.015	0.001	0.035
Lymphocytes	3.233	0.521	0.023	0.218
Mast Cells	3.233	0.048	0.003	0.060
Dendritic Cells	3.233	0.042	0.005	0.075
Macrophages	3.233	0.348	0.023	0.189
OS	3.233	0	0.248	1
OS Time	3.233	bdf	668320.422	581.250
PFI	3.233	0	0.246	1
PFI Time	3.233	423	595954.594	562
gender	0	male	-	-
year_of_birth	0	1940	120.598	16
race	0	white	-	-
ethnicity	0	not hispanic or latino	-	-
year_of_death	0	2011	7.138	4
primary_diagnosis	0	Transitional cell carcinoma	-	-
tumor_stage	0	stage iii	-	-
age_at_diagnosis	0	25297	4975714.935	5728.510
vital_status	0	alive	-	-
morphology	0	8120/3	-	-
days_to_death	0	393	-	-
tissue_or_organ_of_origin	0	Bladder, NOS	-	-
days_to_birth	0	-25297	-	-
site_of_resection_or_biopsy	0	Bladder, NOS	-	-
days_to_last_follow_up	0	232	-	-
cigarettes_per_day	0	0.548	7.961	1.589
weight	0	65	549.958	26.925
bmi	14.319	24.920	41.342	6.643
height	0	170	106.576	14

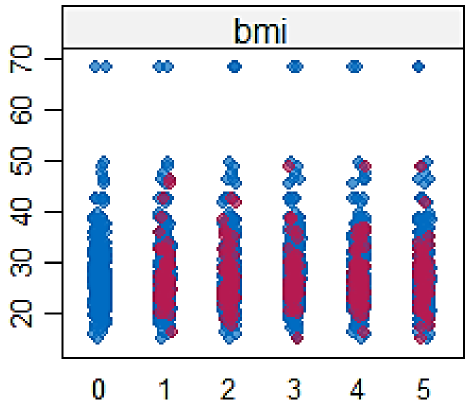
3. Imputation of the *BMI* variable:

As can be seen in the table above, the variable *BMI* presents 14.32% of missing values, therefore, an imputation of this variable will be carried out and later its categorization: *Underweight* (BMI is less than 18.5) - *Normal weight* (BMI is 18.5 to 24.9) - *Overweight* (BMI is 25 to 29.9) - *Obese* (BMI ≥ 30).

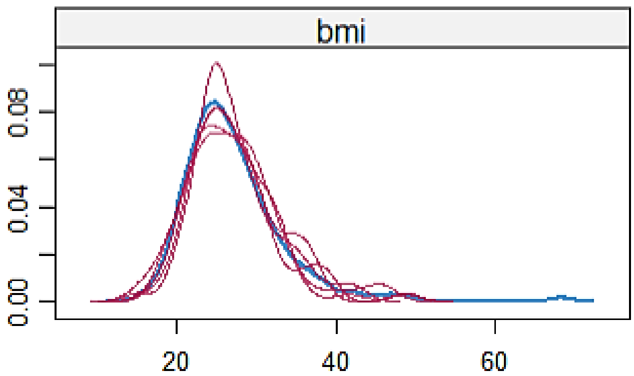
The two methods proposed to carry out the imputation are within the *"mice"* package of RStudio software, they are (1) predictive mean matching *("pmm")* and (2) Bayesian linear regression *("norm")*. These two methods are recommended when working with continuous numerical data:

1) Quality of the Imputations with predictive mean matching:

With the *"stripplot"* function, the distributions of the variables are shown as individual points, the imputed data is shown in purple and the observed data in blue *(Figure 1)*.

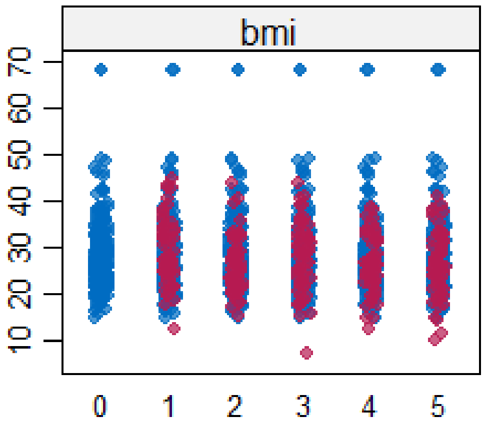


With the *"densityplot"* function, the density of the imputed data is shown, the density for each imputed data set is shown in purple, while the density of the observed data is shown in blue *(Figure 2)*.

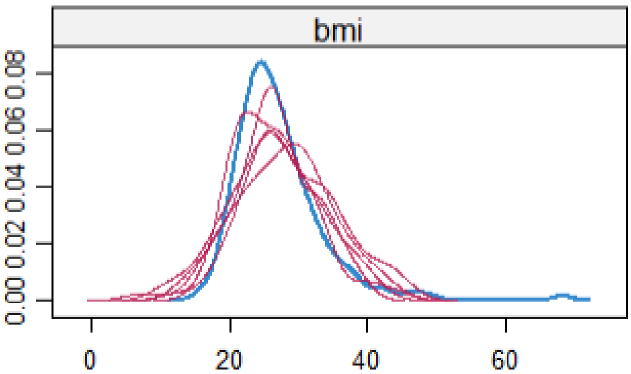


2) Quality of imputations with Bayesian linear regression:

With the *"stripplot"* function *(Figure 3):*



With the *"densityplot"* function *(Figure 4):*



As can be seen, the best results are obtained with the "predictive mean matching" method, so this is the method that will be used to carry out the imputation of the *BMI* variable.

4. Model outputs:

4.1. *Output 1*: Zero Inflated Poisson model, using *Streptococcus parauberis* bacterium and Leukocyte Fraction as covariate:

```
Call:
zeroinfl(formula = fm, data = metadataBCLA_RNA, dist = "poisson", link = "logit")
Count model coefficients (poisson with log link):
                        Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)             -9.8420       0.3804         -25.870      < 2e-16 ***
Leukocyte.Fraction      -3.2005       1.1932         -2.682       0.00731 **

Zero-inflation model coefficients (binomial with logit link):
                        Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)             5.3615        0.9738         5.506        3.67e-08 ***
Leukocyte.Fraction      -2.5434       2.6650         -0.954       0.34
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of iterations in BFGS optimization: 10
Log-likelihood: -44.44 on 4 Df
```

Output 1

4.2. *Output 2:* Zero Inflated Negative Binomial model, using *Streptococcus parauberis* bacterium and Leukocyte Fraction as covariate:

```
Call:
zeroinfl (formula = fm, data = metadataBCLA_RNA, dist = "negbin", link = "logit")
Count model coefficients (negbin with log link):
                        Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)             -7.142        3.633          -1.966       0.0493 *
Leukocyte.Fraction      -16.051       9.117          -1.761       0.0783
Log(theta)              -2.129        3.754          -0.567       0.5706

Zero-inflation model coefficients (binomial with logit link):
                        Estimate      Std.Error      z value      Pr(>|z|)
(Intercept)             5.658         2.228          2.540        0.0111 *
Leukocyte.Fraction      -8.044        6.533          -1.231       0.2182
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta = 0.119
Number of iterations in BFGS optimization: 45
Log-likelihood: -33.37 on 5 Df
```

Output 2

4.3. *Output 3:* Zero Hurdle Poisson Model, using *Streptococcus parauberis* bacterium and Leukocyte Fraction as covariate:

```
Call:
hurdle (formula = fm, data = metadataBCLA_RNA, dist = "poisson")
Count model coefficients (truncated poisson with log link):
                        Estimate    Std.Error    z value    Pr(>|z|)
(Intercept)             -9.8515     0.3793       -25.975    < 2e-16 ***
Leukocyte.Fraction      -3.1601     1.1842       -2.668     0.00762 **

Zero hurdle model coefficients (binomial with logit link):
                        Estimate    Std.Error    z value    Pr(>|z|)
(Intercept)             -5.3305     0.9612       -5.546     2.93e-08 ***
Leukocyte.Fraction      2.3926      2.6161       0.915      0.36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Number of iterations in BFGS optimization: 15
Log-likelihood: -44.5 on 4 Df
```

Output 3

4.4. *Output 4:* Zero Hurdle Negative Binomial Model, using *Streptococcus parauberis* bacterium and Leukocyte Fraction as covariate:

```
Call:
hurdle (formula = fm, data = metadataBCLA_RNA, dist = "negbin")

Count model coefficients (truncated negbin with log link):
                        Estimate    Std.Error    z value    Pr(>|z|)
(Intercept)             -14.17      338.81       -0.042     0.967
Leukocyte.Fraction      -22.50      18.19        -1.237     0.216
Log(theta)              -11.38      338.81       -0.034     0.973

Zero hurdle model coefficients (binomial with logit link):
                        Estimate    Std.Error    z value    Pr(>|z|)
(Intercept)             -5.3305     0.9612       -5.546     2.93e-08 ***
Leukocyte.Fraction      2.3926      2.6161       0.915      0.36
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Theta: count = 0
Number of iterations in BFGS optimization: 354
Log-likelihood: -33.24 on 5 Df
```

Output 4

4.5. *Output 9:* Zero Hurdle Negative Binomial model, using *Streptococcus mitis* bacterium and Tumor Stage as covariate:

Call:
hurdle(formula = fm, data = metadataBCLA_RNA, dist = "negbin")
Count model coefficients (truncated negbin with log link):

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-32.830	68.003	-0.483	0.629
Stage ii	9.575	57.098	0.168	0.867
Stage iii	11.105	57.098	0.194	0.846
Stage iv	9.457	57.098	0.166	0.868
Log(theta)	-11.112	36.938	-0.301	0.764

Zero hurdle model coefficients (binomial with logit link):

	Estimate	Std.Error	z value	Pr(>|z|)
(Intercept)	-2.536e-11	1.000e+00	0.000	1.000
Stage ii	-1.243e+00	1.021e+00	-1.217	0.223
Stage iii	-1.072e+00	1.018e+00	-1.053	0.292
Stage iv	-1.378e+00	1.021e+00	-1.349	0.177

Theta: count = 0
Number of iterations in BFGS optimization: 49

Output 9

4.6. *Output 13: DESeq2* output, using the bacteria without excess zeros and without zeros, and as a covariate, the two-level categorical variable: Sample type,

log2 fold change (MLE)
Wald test p-value
DataFrame with 6 rows and 6 columns

	baseMean	log2FoldChange	lfcSE
	<numeric>	<numeric>	<numeric>
Staphylococcus simulans	78.2910862425755	0.348065025864934	0.830575104714316
Cutibacterium acnes	178.647013593032	4.93359843848002	0.493556953462212
Neisseria gonorrhoeae	42.7774053702636	-2.00758248117571	0.452927974564686
Enterobacter cloacae	46.7175928997971	-2.03207546811005	0.419464389767282
Escherichia coli	36.332787495269	0.48010049093811	0.253699097623582
Klebsiella pneumoniae	71.3703709804922	-1.644371874135	0.313682894232341

	stat	pvalue	padj
	<numeric>	<numeric>	<numeric>
Staphylococcus simulans	0.419065083806785	0.675168567680448	0.675168567680448
Cutibacterium acnes	9.99600634510714	1.5866733140054e-23	9.52003988403239e-23
Neisseria gonorrhoeae	-4.43245415146905	9.31665180530647e-06	1.39749777079597e-05
Enterobacter cloacae	-4.84445287295409	1.26961047660754e-06	2.53922095321508e-06
Escherichia coli	1.89240125579967	0.0584375388462236	0.0701250466154684
Klebsiella pneumoniae	-5.24214709941128	1.58718754237821e-07	4.76156262713463e-07

Output 12